AP • Advanced Placement • College Board

Chuyên luyện thi IELTS/SAT → AP / IB / A-levels

8020 ENGLISH

AP Statistics

Giáo trình luyện thi chuyên sâu

Song ngữ Anh–Việt

8020 ENGLISH

5A/2 Trần Phú, P.4, Q.5, TP.HCM · luyenthi8020.com

Thầy Tường 0332 747 169 · Cô Nancy 0983 422 692

Lời mở đầu · Foreword

Cuốn sách này thuộc bộ giáo trình luyện thi chương trình quốc tế của **8020 English (8020 Education)** — trung tâm luyện thi trọng điểm **IELTS · SAT · AP · IB · A-level · IGCSE** và săn học bổng du học. Toàn bộ nội dung được biên soạn bởi đội ngũ **Giáo sư · Tiến sĩ · Thạc sĩ** tốt nghiệp các trường top đầu thế giới, bám sát khung chương trình chính thức, trình bày đầy đủ lý thuyết, ví dụ mẫu, hình minh hoạ và hệ thống bài tập có lời giải chi tiết.

This book is part of the 8020 English international exam-prep series: complete English theory with Vietnamese explanations, worked examples, illustrations and fully-solved practice, aligned to the official framework.

Thành tích 8020 English

9.0 IELTS cao nhất	**1570** SAT cao nhất
5/5 AP — điểm tuyệt đối	**90/100** IB Diploma
100% GV $\geq$ 8.0 IELTS / 1500 SAT	**CMU** Tiến sĩ ĐH #1 Hoa Kỳ về CS

Đội ngũ tiêu biểu: Thầy Châu Cát Tường (Academic Head, ThS Western Sydney, top 1% IELTS 8.5) · TS Nguyễn Anh Huy (Carnegie Mellon, cựu kỹ sư Amazon) · TS Nguyễn Phước Trường (Toán, ĐH Klagenfurt) · cùng đội ngũ giảng viên SAT 1500–1600, IELTS 8.0–8.5.

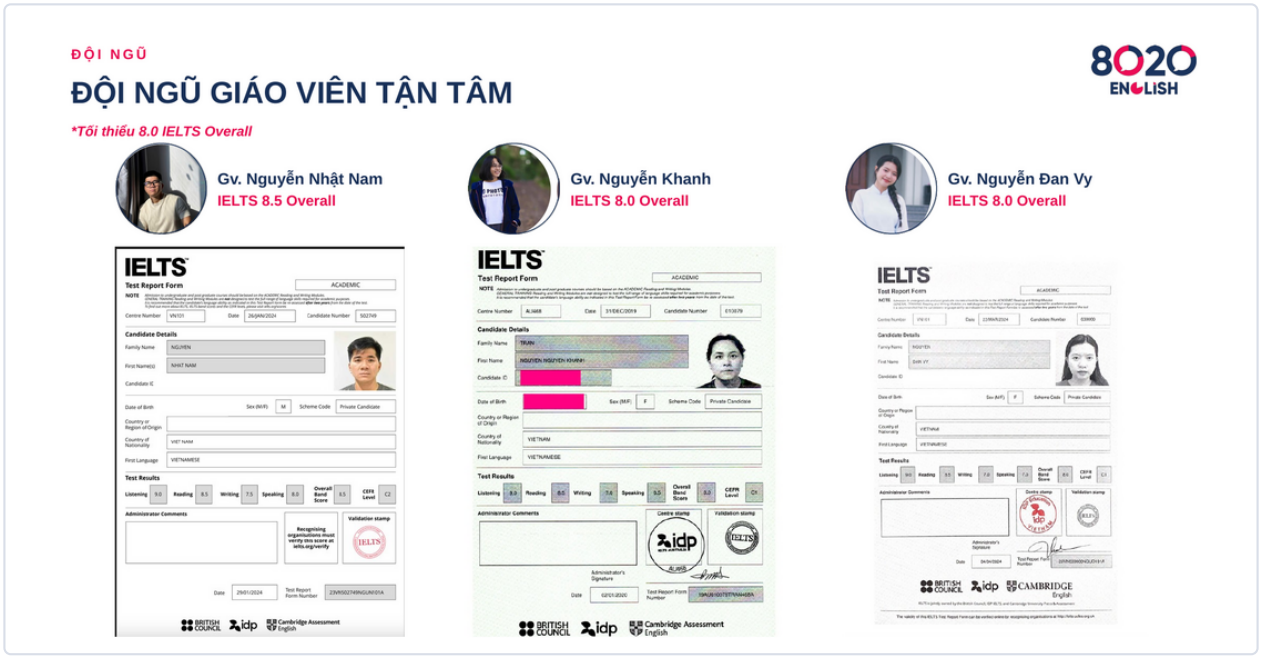
Kết quả học viên — điểm thi thật: IELTS 8.0–9.0 · SAT 1550 / 1570 / 1590 · AP 5/5 (Calculus BC, Microeconomics) · IB 90/100 (Economics) · Duolingo 110 · PTE 90 · TOEFL 120. **Bảng vàng SAT:** Khánh Đăng 1510 · Thảo Nguyên 1520 · Tâm Anh 1530.

"Cuối cùng đã đậu vào trường top như mong muốn." — Phan Thị Thanh Hằng, IELTS Overall 8.5.

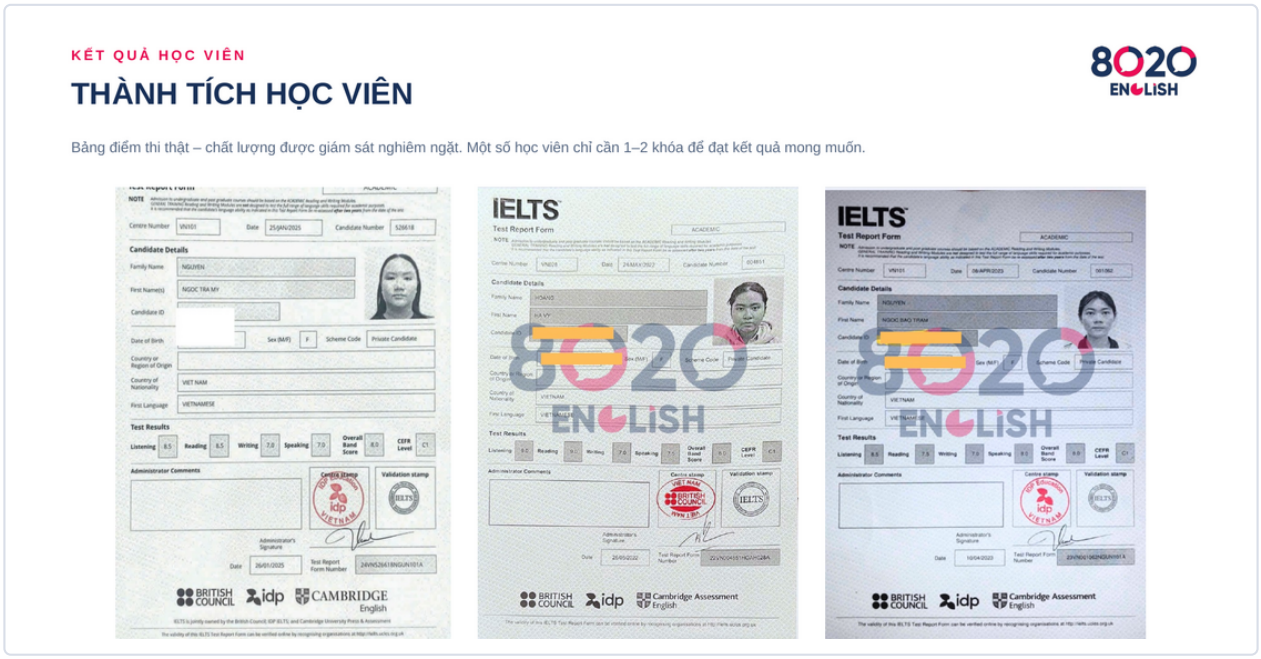
Cách dùng sách hiệu quả: mỗi Unit gồm (1) **Lý thuyết trọng tâm** tiếng Anh kèm **giải thích tiếng Việt**; (2) **Ví dụ mẫu** có lời giải; (3) **Hệ thống bài tập** kèm **lời giải chi tiết**. Hãy tự làm bài trước khi xem lời giải để đạt hiệu quả tối đa.

THÀNH TÍCH THỰC TẾ · 8020 ENGLISH

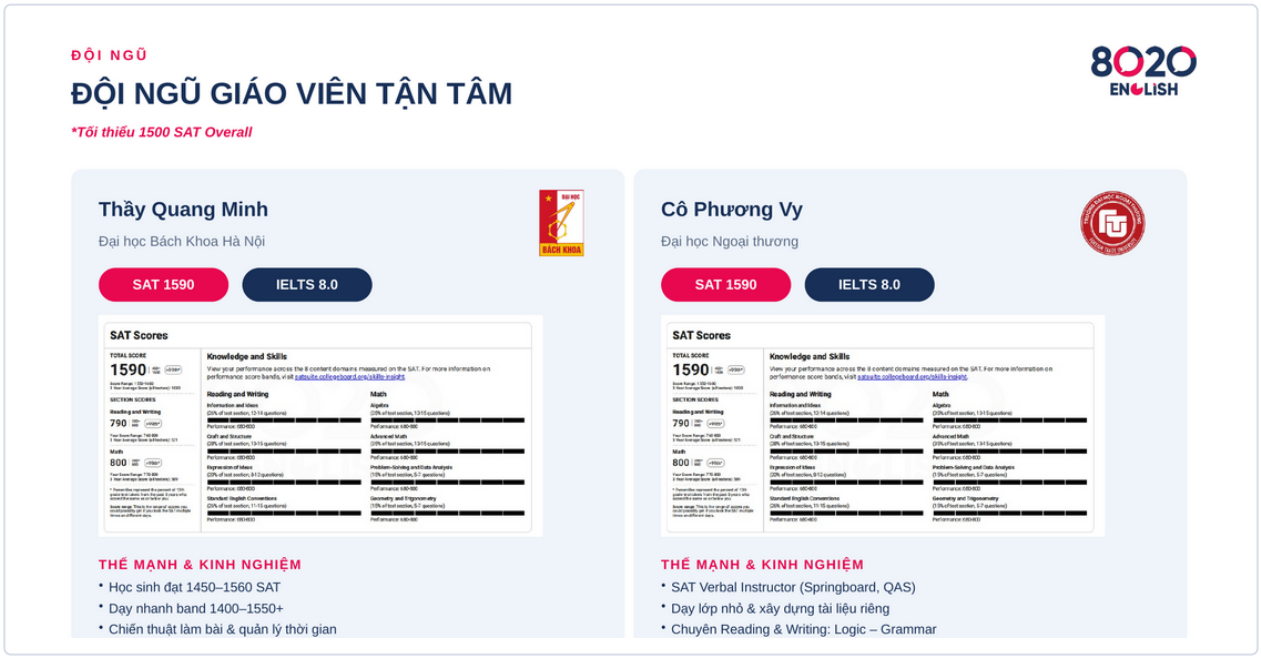
Điểm thi thật · Bằng & bảng điểm thật của giảng viên và học viên 8020 English — **Điểm thật · Thầy thật · Kết quả thật**



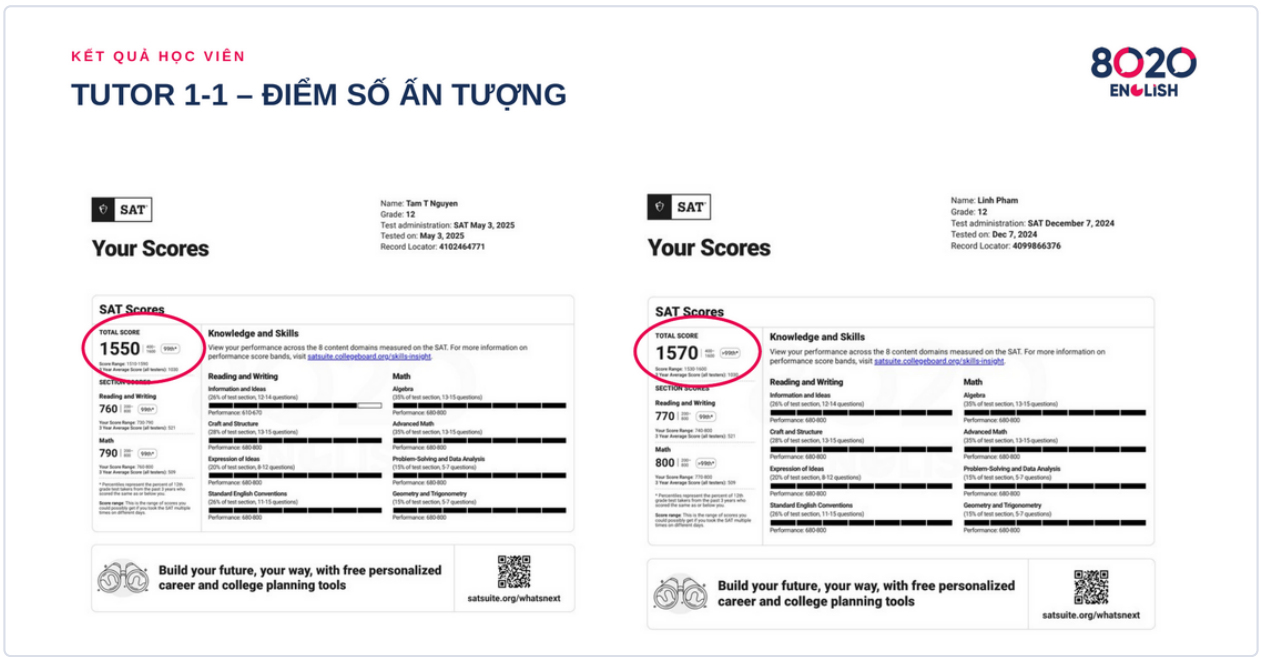
Giảng viên 8020 — IELTS 8.5 & 8.0 Overall (British Council · IDP · Cambridge)



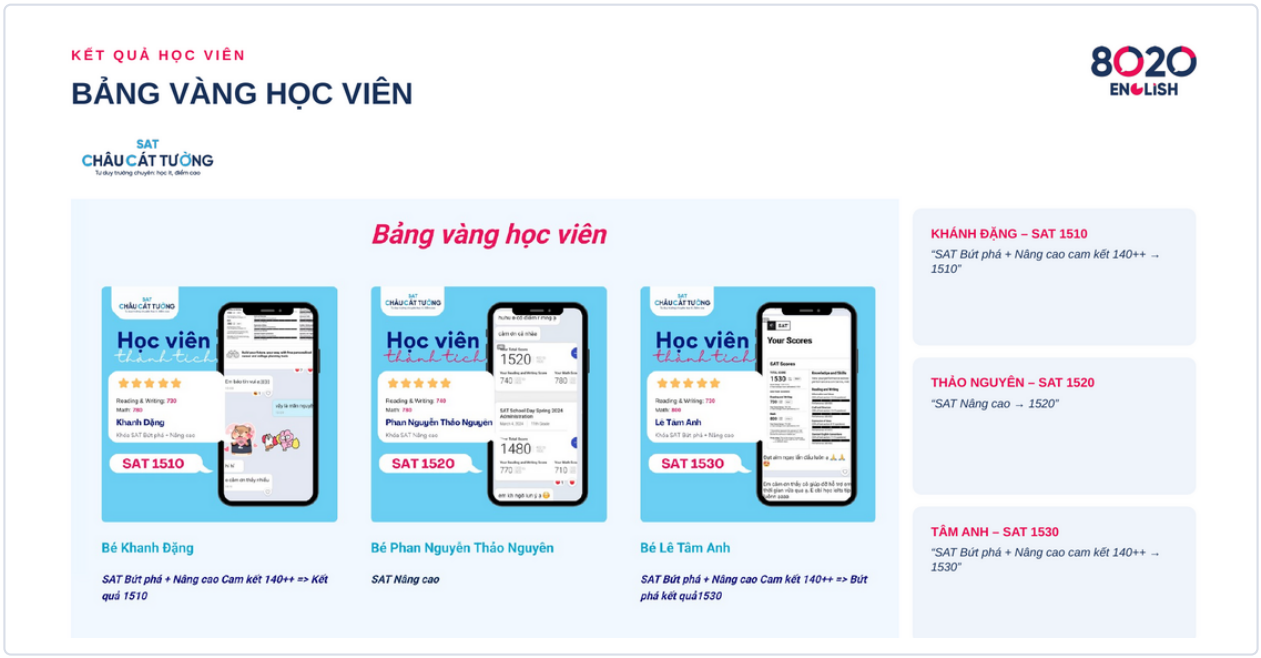
Học viên 8020 — IELTS 8.0 Overall (bảng điểm thật)



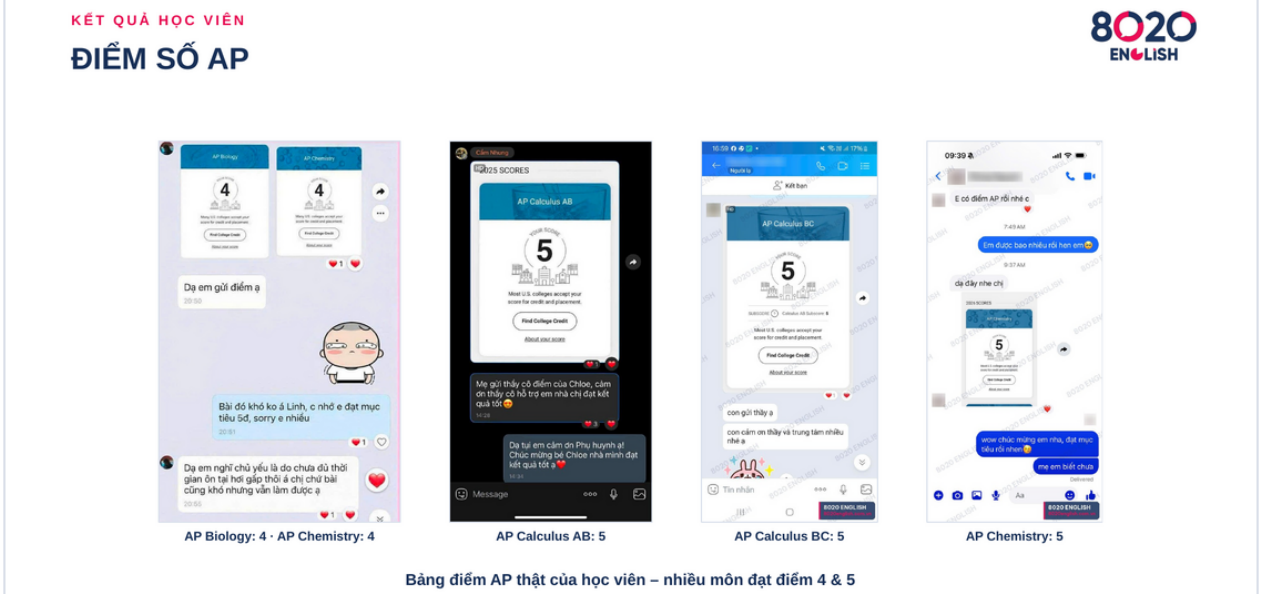
Giảng viên 8020 — SAT 1590 (ĐH Bách Khoa Hà Nội & ĐH Ngoại thương)



Học viên 8020 — SAT 1550 & 1570 (College Board)



Bảng vàng SAT — Học viên đạt 1510 · 1520 · 1530



Điểm AP thật — Calculus AB/BC 5/5 · Biology & Chemistry 4–5 (College Board)

Mục lục · Contents

Exploring One-Variable Data

Mục tiêu Unit: Phân biệt cá thể (individual) và biến (variable), biến định tính và định lượng; lập bảng tần số/tần suất và biểu đồ cột, biểu đồ tròn cho biến định tính; vẽ và đọc dotplot, stemplot (thân–lá), histogram cho biến định lượng; mô tả hình dạng–tâm–độ trải–ngoại lệ (SOCS) trong ngữ cảnh, bao gồm phân phối đều (uniform) và hai đỉnh (bimodal); so sánh hai phân phối bằng back-to-back stemplot hoặc boxplot song song; tính trung bình, trung vị, phương sai, độ lệch chuẩn mẫu s_x và tính đối kháng (resistance) của trung vị; tìm ngũ phân vị (five-number summary), IQR, quy tắc phát hiện ngoại lệ $1.5 \times IQR$, vẽ boxplot có điều chỉnh; hiểu ảnh hưởng của phép biến đổi tuyến tính (cộng hằng số, nhân hằng số) lên tâm và độ trải; nắm vững phân phối chuẩn, quy tắc 68–95–99.7, điểm chuẩn hoá z-score, tính xác suất/phân vị chuẩn bằng bảng chuẩn tắc, bài toán chuẩn ngược (inverse Normal), và đánh giá tính chuẩn qua biểu đồ xác suất chuẩn (Normal probability plot).

Statistics is the science of collecting, organizing, summarizing, and drawing conclusions from data. Before any calculation is possible, every data set must first be organized around two questions: *who* (or what) was measured, and *what* was measured about them. This chapter develops the vocabulary and tools — tables, graphs, and numerical summaries — used to organize and describe a single variable measured on a group of individuals.

1.1 Individuals, Variables, and Distributions

An **individual** is a single object — a person, animal, or thing — described by a set of data. A **variable** is any characteristic of an individual that can take different values for different individuals: height, eye color, favorite subject, and test score are all variables. Every statistical study begins by identifying the individuals under study and the variable(s) recorded for each one.

The **population** is the entire group of individuals that we want information about; a **sample** is the subset of the population that is actually measured or surveyed. A numerical summary of a population (such as the population mean μ or population standard deviation σ) is called a **parameter**; the corresponding summary computed from a sample (such as $\bar{x}$ or s_x) is called a **statistic**. Because collecting data on an entire population is rarely feasible, this chapter's methods are built around describing a *sample* — which is exactly why the standard-deviation formula divides by $n-1$ rather than n.

VN GIẢI THÍCH TIẾNG VIỆT

Tổng thể (population) là toàn bộ nhóm cá thể mà ta muốn tìm hiểu; **mẫu (sample)** là một tập con của tổng thể thực sự được đo/khảo sát. Một con số tóm tắt **tổng thể** (VD μ, σ) gọi là **tham số (parameter)**; con số tóm tắt **mẫu** (VD $\bar{x}$, s_x) gọi là **thống kê (statistic)**. Vì hiếm khi đo được toàn bộ tổng thể, các công thức trong chương này chủ yếu dành cho **mẫu** — đây cũng là lý do độ lệch chuẩn mẫu chia cho $n-1$.

Categorical vs. quantitative variables

A **categorical variable** places an individual into one of several groups or categories (eye color, blood type, favorite subject). Its values are labels, not numbers — even when a label looks

numeric, such as a zip code or a jersey number, it is still categorical because arithmetic on it is meaningless. A **quantitative variable** takes numerical values for which arithmetic operations such as adding, subtracting, and averaging make sense (height in cm, number of siblings, test score). Quantitative variables are further split into **discrete** (a countable list of possible values, such as number of siblings: $0, 1, 2, \ldots$) and **continuous** (any value in an interval, such as height or time).

The **distribution** of a variable tells us what values it takes and how often — the pattern of variation in the data. Distributions of categorical variables are displayed with frequency tables, bar graphs, and pie charts (Section 1.2); distributions of quantitative variables are displayed with dotplots, stemplots, and histograms (Section 1.3), and are described numerically using shape, center, spread, and outliers (Sections 1.3–1.6).

VN **GIẢI THÍCH TIẾNG VIỆT**

Cá thể (individual) là một đối tượng cụ thể được đo hoặc khảo sát; **biến (variable)** là đặc điểm có thể thay đổi giá trị giữa các cá thể. **Biến định tính (categorical)** phân loại cá thể vào các nhóm — giá trị là nhãn, không phải số, kể cả khi nhãn trông giống số (VD mã bưu điện). **Biến định lượng (quantitative)** nhận giá trị số có thể tính toán (cộng, trừ, lấy trung bình), chia thành **rời rạc (discrete)** — đếm được, VD số anh chị em — và **liên tục (continuous)** — VD chiều cao, thời gian. **Phân phối (distribution)** của một biến mô tả các giá trị có thể có và tần suất xuất hiện của từng giá trị đó.

Categorical variables are further described as **nominal** (categories have no natural order — eye color, blood type, favorite subject) or **ordinal** (categories DO have a natural order, but the "distance" between adjacent categories is not necessarily numerically meaningful) — a letter grade of A, B, C, D, F, or a survey response of Poor/Fair/Good/Excellent are both ordinal.

VN **GIẢI THÍCH TIẾNG VIỆT**

Định danh (nominal): các danh mục KHÔNG có thứ tự tự nhiên (VD màu mắt, nhóm máu, môn học yêu thích). **Thứ bậc (ordinal)**: các danh mục CÓ thứ tự tự nhiên nhưng khoảng cách giữa các danh mục liền kề không nhất thiết có ý nghĩa số học (VD điểm chữ A, B, C, D, F, hoặc mức đánh giá Yếu/Trung bình/Khá/Tốt).

Ví dụ mẫu · Classifying individuals and variables

A researcher records, for each of 200 patients at a clinic: patient ID number, age (in years), blood type, resting heart rate (beats per minute), and satisfaction rating (Poor/Fair/Good/Excellent). Classify the individuals, and classify each variable.

The **individuals** are the 200 patients. **Patient ID number** is categorical (nominal) — even though it is written with digits, an ID cannot be meaningfully added, subtracted, or averaged. **Age** is quantitative and discrete if recorded in whole years (technically continuous if measured to great precision). **Blood type** is categorical (nominal) — there is no natural order among A, B, AB, and O. **Resting heart rate** is quantitative and continuous. **Satisfaction rating** is categorical and **ordinal** — Poor $<$ Fair $<$ Good $<$ Excellent has a clear order, but the numerical "gap" between Poor and Fair is not well defined, so averaging these labels as if they were numbers would not be meaningful.

1.2 Representing a Categorical Variable

The distribution of a categorical variable is first organized in a **frequency table**, listing every category together with its **frequency** (count) and its **relative frequency** — the count divided by the total sample size n, usually expressed as a percent.

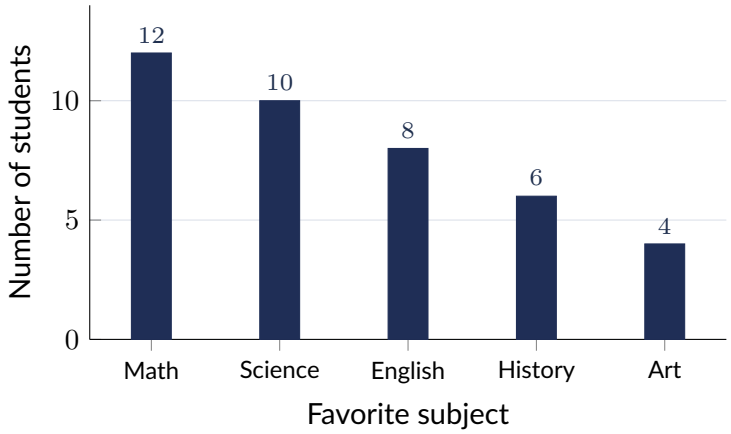
Ví dụ mẫu · Reading a frequency table and bar graph

A teacher surveys 40 students on their favorite subject, obtaining the counts below.

Favorite subject	Frequency	Relative frequency
Math	12	$12/40 = 30\%$
Science	10	$10/40 = 25\%$
English	8	$8/40 = 20\%$
History	6	$6/40 = 15\%$
Art	4	$4/40 = 10\%$
Total	40	100%

Frequency table of favorite subject for a survey of 40 students.

The relative frequencies must sum to 100% (here $30 + 25 + 20 + 15 + 10 = 100$, confirming no arithmetic error). Math is the **mode** (most frequent category), chosen by 30% of students, while Art is the least popular, chosen by only 10%.



Bar graph of favorite subject for the same 40 students (heights = frequencies from the table above).

A **bar graph** displays the count or percent for each category as the height of a separate, equally spaced bar; because every category sits on a common numeric scale, bar graphs make it easy to compare categories precisely and to reorder them by frequency. A **pie chart** instead divides a circle into wedges whose angles are proportional to each category's relative frequency, emphasizing how the categories split up a single whole (100%). Pie charts work well for a small number of categories where "share of the total" is the main message, but grow hard to read — and hard to compare precisely — once there are many categories or several wedges are close in size; a bar graph is almost always easier to read precisely.

VN **GIẢI THÍCH TIẾNG VIỆT**

Bảng tần số (frequency table) liệt kê từng danh mục cùng **tần số (đếm)** và **tần suất tương đối** (tần số chia cho tổng n, đổi ra %) — tổng các tần suất tương đối luôn bằng 100%. **Biểu đồ cột (bar graph)** vẽ mỗi danh mục thành một cột có chiều cao bằng tần số/tần suất trên cùng một trục số, giúp so sánh chính xác giữa các danh mục. **Biểu đồ tròn (pie chart)** chia hình tròn thành các miếng có góc tỉ lệ với tần suất, nhấn mạnh tỉ lệ từng phần so với toàn bộ (100%) nhưng khó

so sánh chính xác khi có nhiều danh mục hoặc các miếng gần bằng nhau.

(!) Lỗi thường gặp: Đừng nhầm **tần số (frequency/count)** với **tần suất tương đối (relative frequency)**: tần số là số lượng cá thể trong một danh mục (một số nguyên, phụ thuộc cỡ mẫu n), còn tần suất tương đối là tỉ lệ phần trăm so với tổng (tần số$/n$), luôn nằm trong khoảng 0% đến 100%. Khi so sánh hai mẫu có cỡ khác nhau (VD 40 học sinh lớp A và 120 học sinh lớp B), phải so sánh bằng **tần suất tương đối**, so sánh trực tiếp tần số (đếm) sẽ gây hiểu lầm.

If a pie chart were drawn instead for the favorite-subject data, each wedge's central angle would be the relative frequency multiplied by $360°$: Math $\to 0.30\times360° = 108°$; Science $\to 0.25\times360° = 90°$; English $\to 0.20\times360° = 72°$; History $\to 0.15\times360° = 54°$; Art $\to 0.10\times360° = 36°$ — and as a check, $108+90+72+54+36 = 360°$ exactly, confirming the wedges account for the full circle.

A **Pareto chart** is a bar graph with categories ordered from most frequent to least frequent (rather than alphabetically or in some arbitrary order), making it immediately clear which categories dominate. Pareto charts are meaningful only for **nominal** categorical variables — for an **ordinal** variable (e.g., letter grades A through F), the natural order of the categories should be preserved in the display instead of being re-sorted by frequency, or the built-in ordering information is lost.

VN **GIẢI THÍCH TIẾNG VIỆT**

Biểu đồ Pareto (Pareto chart) là biểu đồ cột được sắp xếp theo tần số giảm dần (từ danh mục phổ biến nhất đến ít phổ biến nhất), giúp nhận ra ngay danh mục nào chiếm ưu thế — chỉ nên dùng cho biến **định danh (nominal)**. Với biến **thứ bậc (ordinal)**, phải giữ nguyên thứ tự tự nhiên của các danh mục (VD A, B, C, D, F) khi vẽ biểu đồ, không nên sắp xếp lại theo tần số vì sẽ làm mất thông tin thứ tự vốn có.

1.3 Displaying a Quantitative Variable

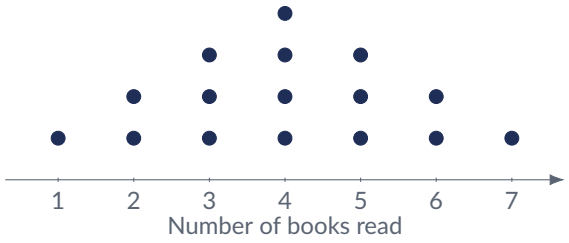
A quantitative variable can be displayed with several different graphs, each with its own strengths: dotplots and stemplots preserve every individual value, while histograms group values into bins to reveal shape more clearly in larger data sets. All three types support the same underlying goal — seeing the overall pattern of variation before summarizing it numerically.

1.3.1 Dotplots

The simplest graph for a quantitative variable is the **dotplot**: draw a horizontal number-line axis, then place one dot above the corresponding value on the axis for each individual, stacking dots when a value repeats. Dotplots preserve every individual data value and make the shape of small-to-moderate data sets easy to see at a glance.

Ví dụ mẫu · Constructing and reading a dotplot

Sixteen students report the number of books they read over summer break: $1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7$.



Dotplot of books read over summer break, $n = 16$ students. Each dot represents one student.

Shape: unimodal (one clear peak, at 4 books) and roughly symmetric — the counts $1, 2, 3, 4, 3, 2, 1$ decrease at the same rate on either side of the peak. **Center:** mean $= \dfrac{1(1) + 2(2) + 3(3) + 4(4) + 5(3) + 6(2) + 7(1)}{16} = \dfrac{64}{16} = 4$ books; median (average of the 8th and 9th sorted values, both 4) $= 4$ books — mean and median agree exactly, consistent with the symmetric shape. **Spread:** values range from 1 to 7 books (range $= 6$). No individual value stands apart from the overall pattern, so there are no obvious outliers.

1.3.2 Stem-and-Leaf Plots (Stemplots)

A **stemplot** (stem-and-leaf plot) splits each data value into a **stem** (all digits except the last) and a **leaf** (the last digit), then lists the leaves for each stem in increasing order. Unlike a histogram, a stemplot keeps every individual data value visible while still displaying the overall shape.

Ví dụ mẫu · Constructing a stemplot

Twenty exam scores (out of 100), sorted: $58, 61, 63, 67, 68, 70, 72, 73, 74, 75, 76, 78, 79, 81, 82, 85, 88, 90, 91, 95$. Using the tens digit as the stem and the units digit as the leaf:

Stem (tens)	Leaves (units)
5	8
6	1 3 7 8
7	0 2 3 4 5 6 8 9
8	1 2 5 8
9	0 1 5

Stemplot of 20 exam scores. Key: $5 \mid 8$ means a score of 58.

Shape: unimodal, with a single peak in the 70s (8 of the 20 scores) and roughly symmetric tails falling off toward the 50s and 90s — no gaps or isolated values suggest outliers. **Reading statistics directly from the display:** because the leaves are already sorted within each stem, the stemplot itself can be used to find the median without re-sorting: with $n = 20$, the median is the average of the 10th and 11th values, which (counting leaves stem by stem: $1 + 4 = 5$ values through stem 6, $5 + 8 = 13$ through stem 7) fall inside the stem-7 row: the 10th value is 75 and the 11th is 76, giving median $= \dfrac{75 + 76}{2} = 75.5$.

Split stems. When many values share the same stem, a stemplot can look too compressed to reveal its shape clearly. **Splitting each stem into two rows** — one for leaves 0–4 and one for leaves 5–9 — spreads the data over roughly twice as many rows and often makes the shape much easier to read, at the cost of a taller display.

Ví dụ mẫu · Split stemplot

Twelve reaction times (tenths of a second): $21, 22, 23, 24, 26, 27, 28, 29, 31, 32, 33, 35$. A single-row stemplot would pile all 8 values from 21 to 29 under one stem, 2. Splitting each stem into a "low" half (leaves 0–4) and a "high" half (leaves 5–9):

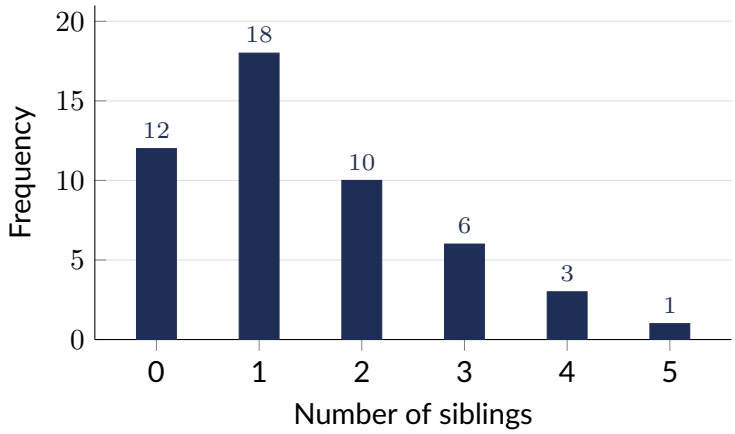
Stem	Leaves
2L (0–4)	1 2 3 4
2H (5–9)	6 7 8 9
3L (0–4)	1 2 3
3H (5–9)	5

Split stemplot of reaction times (tenths of a second), $n = 12$. Key: $2L \mid 1$ means 2.1.

With the split, the data appear roughly evenly spread from 2.1 to 3.3 seconds with a single value trailing off at 3.5 — a pattern that would have been much harder to see packed into just two unsplit stems.

1.3.3 Histograms

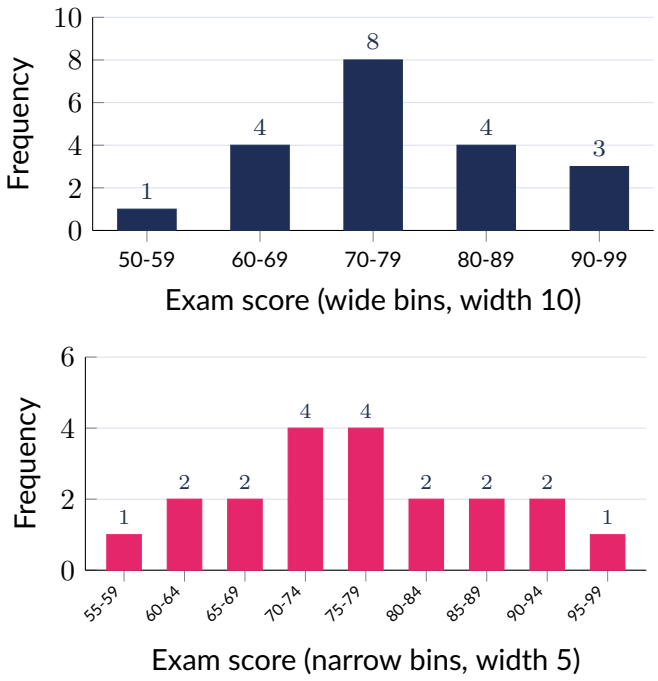
A **histogram** displays a quantitative variable by grouping values into equal-width intervals (**bins**) and drawing a bar over each bin whose height is the frequency (or relative frequency) of values falling in that bin. Unlike a bar graph for categorical data, a histogram's bars touch (since the horizontal axis is a continuous numeric scale) and the order of the bars is fixed by the numeric scale, not chosen for effect.



Frequency histogram of "number of siblings" for 50 students — unimodal, **skewed right** (a long tail toward large values).

This histogram has a single peak at 1 sibling and a long right tail stretching toward 4 and 5 siblings — a shape formalized in the next subsection.

Choosing bin width. A histogram's appearance depends heavily on how many bins are used. Too few, wide bins can hide real structure — even merging two separate peaks into one; too many, narrow bins can make the display noisy and choppy, with many bins holding just 0 or 1 values, which obscures the overall shape instead of revealing it. A useful starting point is somewhere between about 5 and 20 bins, adjusting until the true shape of the data is visible.



The same 20 exam scores from the stemplot example, displayed with wide bins (width 10, top) versus narrow bins (width 5, bottom). Both confirm a single peak in the 70s, but the narrow-bin version reveals additional detail — such as the scores splitting evenly within both the 70s (4 in 70–74, 4 in 75–79) and the 80s (2 and 2) — that the wide-bin version smooths over.

(!) Lỗi thường gặp: Đừng nhầm **histogram** (biến định lượng, trục hoành là thang số liên tục, các cột thường chạm nhau) với **biểu đồ cột (bar graph)** (biến định tính, trục hoành là các danh mục rời rạc, các cột thường tách rời nhau để nhấn mạnh chúng không có thứ tự số học liên tục). Ngoài ra, hình dạng của histogram phụ thuộc vào **độ rộng bin (bin width)**: quá ít bin có thể che mất đặc điểm quan trọng (như hai đỉnh riêng biệt bị gộp thành một), còn quá nhiều bin có thể khiến biểu đồ nhiễu, khó thấy hình dạng tổng thể.

1.3.4 Shape, Center, Spread, and Outliers (SOCS)

Whenever we describe a quantitative distribution, we always discuss four things — collectively remembered as **SOCS**: **shape**, **center** (mean or median), **spread** (range, IQR, or standard deviation), and any **outliers** or unusual features — always stated in the *context* of the data (with correct units and a reference to what was measured), never as bare numbers.

Shape vocabulary. **Symmetric**: the left and right halves are approximate mirror images. **Skewed right**: a long tail extends toward larger values (most data bunched at the low end). **Skewed left**: a long tail extends toward smaller values. **Unimodal**: one clear peak. **Bimodal**: two distinct peaks, often signaling two underlying groups mixed together. **Uniform**: roughly equal frequency across the whole range, with no clear peak at all.

VN **GIẢI THÍCH TIẾNG VIỆT**

Đối xứng (symmetric): hai nửa trái/phải gần như đối xứng nhau. **Lệch phải (skewed right)**: đuôi dài về phía giá trị lớn. **Lệch trái (skewed left)**: đuôi dài về phía giá trị nhỏ. **Một đỉnh (unimodal)**: có một đỉnh rõ ràng. **Hai đỉnh (bimodal)**: có hai đỉnh tách biệt, thường gợi ý dữ liệu là sự trộn lẫn của hai nhóm khác nhau. **Đều (uniform)**: tần số gần như bằng nhau trên toàn miền giá trị, không có đỉnh rõ rệt.

Khái niệm cốt lõi

In a **skewed** distribution, the mean is pulled toward the tail while the median resists it. If mean $>$ median, the distribution is typically skewed right; if mean $<$ median, skewed left. Because of this, the median is the preferred measure of center for skewed distributions or those with outliers.

VN **GIẢI THÍCH TIẾNG VIỆT**

Khi mô tả phân phối một biến định lượng, luôn nêu đủ 4 ý: **hình dạng** (đối xứng, lệch trái/phải), **tâm** (trung bình/trung vị), **độ trải** (khoảng biến thiên/IQR/độ lệch chuẩn) và **ngoại lệ** — và luôn gắn với **ngữ cảnh** đề bài (đơn vị, đối tượng đo). Với phân phối lệch, trung vị ổn định hơn (resistant) trung bình vì trung bình bị "kéo" về phía đuôi dài.

Outliers: causes and handling

An outlier can arise from a genuine, if unusual, individual (a true extreme value that legitimately belongs in the data set) or from an **error** — a mis-measurement, a data-entry mistake, or an instrument malfunction. Before deciding how to treat an outlier, it is good practice to investigate its cause whenever possible. A confirmed data-entry error should be corrected or removed; a genuine extreme observation should normally be *kept* but reported carefully — often alongside a summary computed both with and without it, since (as seen throughout this chapter) an outlier can dramatically change the mean, range, and standard deviation while barely touching the median and IQR.

VN **GIẢI THÍCH TIẾNG VIỆT**

Một ngoại lệ (outlier) có thể là một cá thể **thực sự** bất thường (hợp lệ, chỉ là hiếm gặp) hoặc là kết quả của **sai số** — đo sai, nhập liệu sai, hoặc thiết bị lỗi. Trước khi quyết định xử lý ra sao, nên tìm hiểu nguyên nhân nếu có thể: nếu là lỗi nhập liệu, nên sửa hoặc loại bỏ; nếu là giá trị cực đoan hợp lệ, thường nên **giữ lại** nhưng báo cáo cẩn thận — có thể trình bày cả hai bản tóm tắt (có và không có ngoại lệ), vì ngoại lệ có thể làm thay đổi mạnh trung bình, range, và độ lệch chuẩn trong khi hầu như không ảnh hưởng đến trung vị và IQR.

1.3.5 Comparing Two Distributions

A required skill is comparing two quantitative distributions using SOCS — never just describing each one separately, but explicitly comparing shape, center, and spread *between* the groups, in context. A **back-to-back stemplot** — one shared stem column with leaves for group A on the left (read outward from the stem) and leaves for group B on the right — is a convenient way to compare two distributions directly.

Ví dụ mẫu · Comparing two distributions with a back-to-back stemplot

Ten students in Class A and ten in Class B take the same exam:
Class A: $62, 65, 68, 71, 73, 75, 78, 82, 85, 90$. Class B: $55, 58, 63, 66, 70, 72, 74, 77, 80, 83$.

Class A (leaves)	Stem	Class B (leaves)
	5	5 8
8 5 2	6	3 6
8 5 3 1	7	0 2 4 7
5 2	8	0 3
0	9	

Back-to-back stemplot comparing Class A and Class B exam scores. Key: $6\,|\,2$ on the Class A side means 62; $7\,|\,0$ on the Class B side means 70.

Shape: both distributions are roughly unimodal, without gaps or obvious outliers.

Center: Class A: mean $= \dfrac{749}{10} = 74.9$, median $= \dfrac{73+75}{2} = 74$. Class B: mean $= \dfrac{698}{10} = 69.8$, median $= \dfrac{70+72}{2} = 71$. Class A's scores are centered noticeably higher than Class B's (about 5 points higher on both mean and median) — Class A tended to score better.

Spread: using the exclude-the-median method ($n = 10$, even, split into two halves of 5): Class A has $Q_1 = 68$, $Q_3 = 82$, so $IQR_A = 14$; Class B has $Q_1 = 63$, $Q_3 = 77$, so $IQR_B = 14$ — the two classes have essentially the *same* spread of scores, even though their centers differ. Ranges also match: $90 - 62 = 28$ for A, $83 - 55 = 28$ for B.

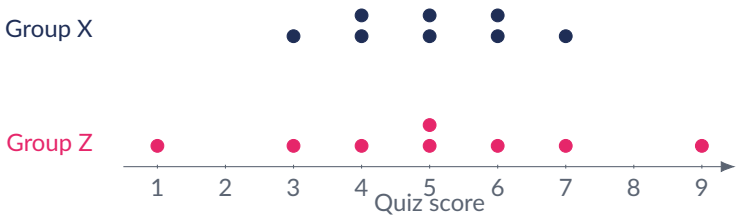
Outliers: for Class A, the fences are $68 - 1.5(14) = 47$ and $82 + 1.5(14) = 103$ — every value (62 to 90) lies inside, so no outliers. For Class B, the fences are $63 - 1.5(14) = 42$ and $77 + 1.5(14) = 98$ — again every value lies inside, so no outliers either.

Conclusion in context: Class A's exam scores are centered about 5 points higher than Class B's, but the two classes show essentially the same amount of variability (equal IQR and range), and neither class has any outlying scores.

Parallel **dotplots** — two dotplots stacked on the same numeric axis — offer a second convenient way to compare distributions, especially useful when every individual value should stay visible.

Ví dụ mẫu · Comparing distributions with parallel dotplots

Two groups of 8 students take the same short quiz (scored 0 to 10):
Group X: $3, 4, 4, 5, 5, 6, 6, 7$. Group Z: $1, 3, 4, 5, 5, 6, 7, 9$.



Parallel dotplots comparing Group X (navy, top) and Group Z (pink, bottom) quiz scores.

Center: Group X has mean $= \dfrac{40}{8} = 5$ and median $= 5$. Group Z has mean $= \dfrac{40}{8} = 5$ and median $= 5$ as well — the two groups are centered at *exactly the same* value.

Spread: Group X: lower half $\{3, 4, 4, 5\} \Rightarrow Q_1 = 4$; upper half $\{5, 6, 6, 7\} \Rightarrow Q_3 = 6$; $IQR_X = 2$; range $= 7 - 3 = 4$. Group Z: lower half $\{1, 3, 4, 5\} \Rightarrow Q_1 = 3.5$; upper half $\{5, 6, 7, 9\} \Rightarrow Q_3 = 6.5$; $IQR_Z = 3$; range $= 9 - 1 = 8$ — *twice* the range of Group X.

Shape and outliers: both dotplots are roughly mound-shaped, though Group Z's is noticeably flatter and more spread out. Checking Group Z's extremes against the $1.5 \times IQR$ rule: upper fence $= 6.5 + 1.5(3) = 11$ and lower fence $= 3.5 - 1.5(3) = -1$; both 9 and 1 lie inside these fences, so neither group has a flagged outlier.

Conclusion in context: the two groups performed, on average, identically (same mean and median quiz score), but Group Z's scores were considerably more variable ($IQR = 3$, range $= 8$) than Group X's ($IQR = 2$, range $= 4$) — Group X's students performed more consistently with one another, even though the two groups' typical (central) performance was the same.

(!) Lỗi thường gặp: Khi so sánh hai phân phối, tránh viết mơ hồ kiểu "nhóm A có nhiều dữ liệu hơn" hoặc "hai nhóm khác nhau" — phải nêu RÕ theo từng ý của SOCS: so sánh **tâm** (trung bình/trung vị nhóm nào cao hơn, chênh lệch bao nhiêu, đúng đơn vị), so sánh **độ trải** (IQR hoặc độ lệch chuẩn nhóm nào lớn hơn), và nêu rõ **ngoại lệ** nếu có — luôn gắn với ngữ cảnh của đề bài, không chỉ nói suông là "khác nhau" mà không định lượng.

1.4 Measuring Center and Spread

Graphs reveal shape, but precise comparisons require numbers. This section develops the two most common numerical summaries of center (mean and median) and the most common numerical summary of spread built around the mean (standard deviation), together with the crucial idea of **resistance** — how sensitive a statistic is to unusual values.

1.4.1 Mean and Standard Deviation

Mean: $\bar{x} = \dfrac{\sum x_i}{n}$. **Sample variance:** $s_x^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$. **Sample standard deviation:** $s_x = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}$.

The standard deviation measures the typical (roughly average) distance of a value from the mean; it is measured in the *same units* as the original data (variance is in *squared* units, which is why s_x is reported far more often than s_x^2).

VN **GIẢI THÍCH TIẾNG VIỆT**

Trung bình ($\bar{x}$) là tổng các giá trị chia cho số lượng n. **Phương sai mẫu (s_x^2)** và **độ lệch chuẩn mẫu (s_x)** đo độ trải: lấy từng độ lệch so với trung bình, bình phương lên (để loại dấu âm), cộng lại, rồi chia cho $n-1$ (không phải n) — độ lệch chuẩn là căn bậc hai của phương sai, có cùng đơn vị với dữ liệu gốc nên dễ diễn giải hơn phương sai.

Ví dụ mẫu · Sample variance and standard deviation

Find s_x^2 and s_x for the data $2, 4, 4, 4, 5, 5, 7, 9$.

$\bar{x} = \dfrac{2+4+4+4+5+5+7+9}{8} = \dfrac{40}{8} = 5$. Deviations from the mean: $-3, -1, -1, -1, 0, 0, 2, 4$; squared: $9, 1, 1, 1, 0, 0, 4, 16$, which sum to 32.

$$s_x^2 = \frac{32}{8-1} = \frac{32}{7} \approx 4.57, \qquad s_x = \sqrt{\frac{32}{7}} \approx \sqrt{4.571} \approx 2.14.$$

Note the divisor is $n-1 = 7$, *not* n — this is what makes s_x a sample statistic. The variance (≈ 4.57 "squared units") is rarely reported directly; the standard deviation (≈ 2.14, in the original units) is the number we interpret and compare.

Why $n-1$? Dividing by $n-1$ instead of n — sometimes called **Bessel's correction** — produces a better estimate of the true (unknown) population variance σ^2. Deviations are computed from the *sample's own* mean $\bar{x}$ rather than the true population mean μ, and $\bar{x}$ is, by construction, the value that makes the sum of squared deviations as small as possible for that particular sample — so using $\bar{x}$ tends to make the squared deviations, on average, slightly *too small*. Dividing by the smaller number $n-1$ compensates for this, so that s_x^2 is, on average across many samples, an accurate (unbiased) estimate of σ^2.

Ví dụ mẫu · Mean from a frequency table

Twenty families report their number of pets: 0 pets (5 families), 1 pet (8 families), 2 pets (4 families), 3 pets (3 families). Find the mean number of pets per family.

Each value is weighted by its frequency:

$$\bar{x} = \frac{0(5) + 1(8) + 2(4) + 3(3)}{5 + 8 + 4 + 3} = \frac{0 + 8 + 8 + 9}{20} = \frac{25}{20} = 1.25 \text{ pets.}$$

This weighted-average approach — multiplying each distinct value by its frequency, summing, and dividing by the total count — gives exactly the same answer as adding up all 20 raw values individually and dividing by 20, but is far faster whenever data are already organized in a frequency table.

1.4.2 Resistance: Mean versus Median

A statistic is called **resistant** if it is not strongly affected by outliers or by changing the shape of the tail of a distribution. The **median** is resistant — it depends only on the position of the middle value(s), not their exact size. The **mean** is *not* resistant, since every value (including any extreme one) enters the sum directly.

Ví dụ mẫu · Resistance of the median

Five employees earn (in thousands of dollars): $40, 42, 45, 48, 50$. Mean $= \frac{225}{5} = 45$; median $= 45$ (the middle value) — mean and median agree.

Now suppose the highest earner is actually the CEO, earning $\$500{,}000$ instead of $\$50{,}000$: the data become $40, 42, 45, 48, 500$. New mean $= \frac{675}{5} = 135$ — more than *triple* the original mean, and higher than 4 of the 5 salaries. New median (still the middle of the sorted list $40, 42, 45, 48, 500$) $= 45$ — completely **unchanged**.

The median still describes a "typical" employee's salary; the mean, dragged upward by a single outlier, no longer does. This is exactly why salary and income data (almost always right-skewed, with a few very high earners) are usually summarized by the **median**, not the mean.

Statistic	Measures	Resistant to outliers?
Mean $\bar{x}$	center	No
Median	center	Yes
Range	spread	No
Interquartile range (IQR)	spread	Yes
Standard deviation s_x	spread	No

Resistance of common summary statistics. Resistant statistics (median, IQR) depend only on the position of the middle values; non-resistant statistics (mean, range, standard deviation) are pulled by the exact size of every value, including extreme ones.

Because outliers and skewness affect the mean and the standard deviation far more than the median and IQR, a distribution's **shape** should always be checked (via a graph) *before* deciding which pair of summary statistics to report: mean and standard deviation for roughly symmetric data with no outliers, or median and IQR for skewed data or data containing outliers.

1.5 The Five-Number Summary, IQR, and Boxplots

Because the mean and standard deviation are pulled by outliers and skewness, statisticians also rely on an alternative, resistant numerical summary built entirely from the data's order — the five-number summary — together with its standard graphical display, the boxplot.

1.5.1 Five-Number Summary and the Outlier Rule

Five-number summary: minimum, Q_1, median, Q_3, maximum. $IQR = Q_3 - Q_1$ measures the spread of the middle 50% of the data. **Range** = maximum − minimum.

Outlier rule: a value is a suspected outlier if it is below $Q_1 - 1.5 \cdot IQR$ (the **lower fence**) or above $Q_3 + 1.5 \cdot IQR$ (the **upper fence**).

VN **GIẢI THÍCH TIẾNG VIỆT**

Ngũ phân vị (five-number summary) gồm giá trị nhỏ nhất, Q_1, trung vị, Q_3, giá trị lớn nhất. **IQR** $= Q_3 - Q_1$ đo độ trải của 50% dữ liệu ở giữa — ổn định hơn nhiều so với **khoảng biến thiên (range)** = giá trị lớn nhất − nhỏ nhất, vì range chỉ dựa vào đúng 2 giá trị cực trị. **Quy tắc phát hiện ngoại lệ**: một giá trị bị coi là ngoại lệ nghi ngờ nếu nằm dưới $Q_1 - 1.5 \times IQR$ hoặc trên $Q_3 + 1.5 \times IQR$.

Ví dụ mẫu · Five-number summary & outliers

Eleven quiz scores, already sorted: $12, 14, 15, 15, 16, 18, 18, 19, 20, 21, 35$. Find the five-number summary and check for outliers.

$n = 11$ is odd, so exclude the median from both halves. Median (6th value) $= 18$. Lower half $\{12, 14, 15, 15, 16\} \Rightarrow Q_1 = 15$. Upper half $\{18, 19, 20, 21, 35\} \Rightarrow Q_3 = 20$. So $IQR = 20 - 15 = 5$.

$$\text{Upper fence} = Q_3 + 1.5(IQR) = 20 + 7.5 = 27.5.$$

Since $35 > 27.5$, the value 35 is a **suspected outlier**; the lower fence is $15 - 7.5 = 7.5$ and no value is below it.

Some treatments further distinguish **mild outliers** (beyond $1.5 \times IQR$ but not beyond $3 \times IQR$) from **extreme outliers** (beyond $3 \times IQR$ from Q_1 or Q_3). Applying this to the quiz-score example above: the extreme upper fence is $Q_3 + 3(IQR) = 20 + 3(5) = 35$. Because 35 is not *strictly greater than* 35, it is classified as a **mild** outlier only, not an extreme one — a reminder to apply fence inequalities carefully at exact boundary values.

VN **GIẢI THÍCH TIẾNG VIỆT**

Một số tài liệu còn phân biệt **ngoại lệ nhẹ (mild outlier)** — vượt quá $1.5 \times IQR$ nhưng chưa tới $3 \times IQR$ — với **ngoại lệ mạnh (extreme outlier)** — vượt quá $3 \times IQR$. Cần chú ý dấu bất đẳng thức nghiêm ngặt "$>$": nếu một giá trị đúng bằng ranh giới (fence), giá trị đó KHÔNG bị coi là ngoại lệ ở mức đó.

1.5.2 Boxplots

A **modified boxplot** plots the five-number summary directly: a box from Q_1 to Q_3 (with a line at the median), whiskers extending out to the smallest and largest values that are *not* flagged as outliers, and any outliers marked individually as separate points beyond the whiskers.

Modified boxplot of the quiz scores: box $[Q_1, Q_3] = [15, 20]$, median $= 18$, whiskers to 12 and 21, outlier at 35 shown as a dot.

1.5.3 Parallel Boxplots: Comparing Groups

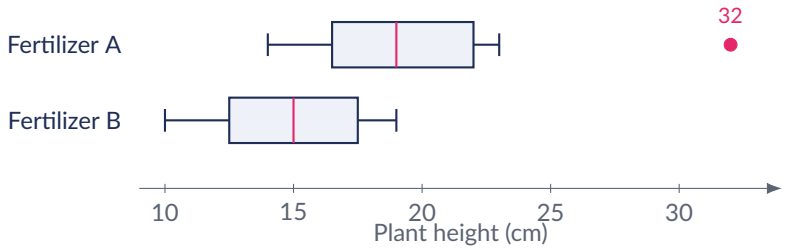
Placing two or more boxplots on the *same* numeric axis — **parallel boxplots** — is a compact, standard way to compare several distributions' centers, spreads, and outliers at once.

Ví dụ mẫu · Parallel boxplots comparing two groups

Plant heights (cm) after eight weeks are recorded for nine plants under Fertilizer A and nine under Fertilizer B:
Fertilizer A: $14, 16, 17, 18, 19, 20, 21, 23, 32$. Fertilizer B: $10, 12, 13, 14, 15, 16, 17, 18, 19$.

Fertilizer A ($n = 9$, odd — exclude median from both halves): median (5th value) $= 19$; lower half $\{14, 16, 17, 18\} \Rightarrow Q_1 = 16.5$; upper half $\{20, 21, 23, 32\} \Rightarrow Q_3 = 22$; $IQR_A = 22 - 16.5 = 5.5$. Upper fence $= 22 + 1.5(5.5) = 30.25$; since $32 > 30.25$, the value 32 is an **outlier**, so the upper whisker stops at the largest non-outlier value, 23.

Fertilizer B ($n = 9$): median $= 15$; lower half $\{10, 12, 13, 14\} \Rightarrow Q_1 = 12.5$; upper half $\{16, 17, 18, 19\} \Rightarrow Q_3 = 17.5$; $IQR_B = 17.5 - 12.5 = 5.0$. Upper fence $= 17.5 + 1.5(5.0) = 25$; the maximum (19) is well below this, so there are **no outliers** in Fertilizer B.



Parallel modified boxplots comparing plant height (cm) under Fertilizer A (median 19, $IQR = 5.5$, one high outlier at 32) and Fertilizer B (median 15, $IQR = 5.0$, no outliers).

Comparison in context: the median height under Fertilizer A (19 cm) is 4 cm higher than under Fertilizer B (15 cm), and the two treatments have similar spread ($IQR_A = 5.5$ vs. $IQR_B = 5.0$ cm). Fertilizer A produced one unusually tall plant (32 cm, an outlier), while Fertilizer B produced no outliers.

Extending to three or more groups. Parallel boxplots generalize directly to three (or more) groups — simply add another row to the same shared axis. Suppose a third treatment, Fertilizer C, is also tested on nine plants: $15, 16, 16, 17, 17, 18, 18, 19, 20$. Median (5th of 9) $= 17$; lower half $\{15, 16, 16, 17\} \Rightarrow Q_1 = 16$; upper half $\{18, 18, 19, 20\} \Rightarrow Q_3 = 18.5$; $IQR_C = 18.5 - 16 = 2.5$. Fences: upper $= 18.5 + 1.5(2.5) = 22.25$ and lower $= 16 - 1.5(2.5) = 12.25$ — every value (15 to 20) lies inside, so Fertilizer C has **no outliers**.

Comparing all three treatments: Fertilizer A has the highest median (19 cm) but also the largest spread ($IQR = 5.5$) and one outlier; Fertilizer B has the lowest median (15 cm) with moderate spread ($IQR = 5.0$); Fertilizer C's median (17 cm) falls between the other two, but its spread ($IQR = 2.5$) is by far the smallest of the three groups — Fertilizer C produced the most **consistent** growth, even though it was not the tallest on average. A three-row parallel boxplot, adding one more row to the figure above on the same numeric axis, would display all three comparisons at a glance.

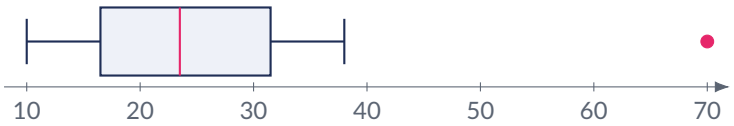
Ví dụ mẫu · Bringing it together: a full numerical summary

Commute times (minutes) for 12 employees at a company, sorted: $10, 12, 15, 18, 20, 22, 25, 28, 30, 33, 38, 70$. Give a complete numerical and graphical summary.

Five-number summary ($n = 12$, even): median $= \dfrac{22+25}{2} = 23.5$; lower half $\{10, 12, 15, 18, 20, 22\} \Rightarrow Q_1 = \dfrac{15+18}{2} = 16.5$; upper half $\{25, 28, 30, 33, 38, 70\} \Rightarrow Q_3 = \dfrac{30+33}{2} = 31.5$; minimum $= 10$, maximum $= 70$. So $IQR = 31.5 - 16.5 = 15$.

Outliers: upper fence $= 31.5 + 1.5(15) = 54$; since $70 > 54$, the value 70 is a suspected outlier — and since the extreme fence $31.5 + 3(15) = 76.5$ is still above 70, this is a *mild*, not extreme, outlier. Lower fence $= 16.5 - 22.5 = -6$; the minimum (10) is not below it.

Center: mean $= \dfrac{321}{12} = 26.75$ minutes; median $= 23.5$ minutes. Mean $>$ median, consistent with the single high outlier pulling the mean upward.



Modified boxplot of the commute-time data: box $[Q_1, Q_3] = [16.5, 31.5]$, median $= 23.5$, whiskers to 10 and 38, outlier at 70.

Full description in context: the distribution of commute times is **right-skewed** (mean pulled above the median) with one high **outlier** at 70 minutes; setting that outlier aside, the middle half of employees commute between 16.5 and 31.5 minutes, with a typical (median) commute of about 23.5 minutes.

1.6 Effect of Linear Transformations on Center and Spread

Data are often **transformed** — rescaled or shifted — for example, converting units, curving a test, or correcting for inflation. A **linear transformation** has the form $y = a + bx$, and its effect on summary statistics follows two clean rules.

Adding a constant a to every value ($y = x + a$): shifts the mean, median, minimum, and maximum all by a — but leaves the **range, IQR, and standard deviation unchanged** (spread and shape are unaffected by a pure shift).

Multiplying every value by a positive constant b ($y = bx$): scales the mean, median, minimum, maximum, range, IQR, *and* standard deviation **all by the factor** b (shape unchanged). If $b < 0$, positions are also reflected, and the standard deviation is scaled by $|b|$, since a standard deviation can never be negative.

VN **GIẢI THÍCH TIẾNG VIỆT**

Cộng hằng số vào mọi giá trị chỉ **dịch chuyển** tâm (trung bình, trung vị, min, max) theo đúng hằng số đó, còn **độ trải** (range, IQR, độ lệch chuẩn) hoàn toàn KHÔNG đổi. **Nhân với hằng số dương** b thì tâm VÀ độ trải đều được **nhân lên theo hệ số** b — cả hai loại thống kê đều thay đổi theo cùng tỉ lệ. Nếu nhân với số âm, phải lấy giá trị tuyệt đối $|b|$ cho độ lệch chuẩn (vì độ lệch chuẩn không âm).

Ví dụ mẫu · Applying the transformation rules

A short quiz (out of a low maximum) is given to four students: $20, 24, 26, 30$. Direct computation gives mean $= 25$, median $= 25$, range $= 10$, $IQR = 6$ (using $Q_1 = 22$, $Q_3 = 28$), and $s_x \approx 4.16$ (variance $= 52/3 \approx 17.33$).

(a) Curve: add 5 points to every score. New data: $25, 29, 31, 35$. New mean $= 25 + 5 = 30$; new median $= 25 + 5 = 30$; range $= 35 - 25 = 10$ (**unchanged**); new $Q_1 = 27$, $Q_3 = 33 \Rightarrow IQR = 6$ (**unchanged**); the deviations from the new mean are identical to before $(-5, -1, 1, 5)$, so $s_x \approx 4.16$ (**unchanged**).

(b) Rescale to a 100-point scale: multiply every original score by 2. New data: $40, 48, 52, 60$. New mean $= 25 \times 2 = 50$; new median $= 25 \times 2 = 50$; range $= 10 \times 2 = 20$; new $Q_1 = 44$, $Q_3 = 56 \Rightarrow IQR = 6 \times 2 = 12$; new $s_x^2 = \dfrac{100 + 4 + 4 + 100}{3} = \dfrac{208}{3} \approx 69.33 \Rightarrow s_x \approx 8.33 = 4.16 \times 2$.

This confirms both rules exactly: adding shifted only the centers and extremes, leaving every spread measure fixed; multiplying scaled *every* statistic (center and spread alike) by the same factor of 2.

(!) Lỗi thường gặp: Nhiều học sinh nghĩ rằng cộng một hằng số vào mọi giá trị sẽ làm thay đổi **độ lệch chuẩn (standard deviation)** hoặc **IQR** — điều này SAI: phép **cộng (adding a constant)** chỉ dịch chuyển tâm của phân phối, không làm thay đổi độ trải. Chỉ phép **nhân (multiplying by a constant)** mới làm co giãn độ trải theo đúng hệ số đó (dùng giá trị tuyệt đối nếu hệ số âm, vì độ lệch chuẩn không thể âm).

1.6.1 Standardizing as a Special Linear Transformation

The z-score formula introduced later in this chapter, $z = \dfrac{x - \mu}{\sigma}$, is itself a linear transformation: it can be rewritten as $z = \left(-\dfrac{\mu}{\sigma}\right) + \left(\dfrac{1}{\sigma}\right) x$, i.e., *subtracting* the constant μ (shifting the mean to 0) and then *multiplying* by the constant $\dfrac{1}{\sigma}$ (scaling the standard deviation to exactly 1).

Applying the two transformation rules in sequence: subtracting μ shifts the mean from μ to 0 without changing the spread; multiplying by $\dfrac{1}{\sigma}$ then scales that spread from σ down to exactly 1. **Standardized data (z-scores) therefore always have mean 0 and standard deviation 1**, regardless of the shape or original units of the data — a fact that underlies the Normal-distribution work in the rest of this chapter and reappears throughout the course (e.g., correlation and least-squares regression in Unit 2).

VN **GIẢI THÍCH TIẾNG VIỆT**

Công thức $z = \dfrac{x - \mu}{\sigma}$ thực chất là một phép **biến đổi tuyến tính**: **trừ** μ (dịch trung bình về 0, không đổi độ trải) rồi **nhân** với $\dfrac{1}{\sigma}$ (co giãn độ lệch chuẩn về đúng 1). Vì vậy, sau khi chuẩn hoá (standardize) BẤT KỲ tập dữ liệu nào — không nhất thiết phải phân phối chuẩn — dữ liệu mới luôn có trung bình $= 0$ và độ lệch chuẩn $= 1$.

1.7 The Normal Distribution

Many quantitative variables encountered in practice — physical measurements, standardized test scores, manufactured-part dimensions — follow a strikingly similar bell-shaped pattern. This section introduces the mathematical model for that pattern, the **Normal distribution**, and the tools used to compute probabilities and percentiles from it.

1.7.1 Density Curves and the Normal Model

A **density curve** is a smooth curve that models the overall shape of a distribution: it is always on or above the horizontal axis, and the total area underneath it equals exactly 1 (representing 100% of the data). The area under a density curve above any interval gives the proportion of observations in that interval. A density curve can be thought of as the limiting shape approached by a histogram of the same variable as the sample size grows very large and the bin width shrinks toward zero — the jagged bars smooth out into a continuous curve, while the total area (now 1 instead of the total count) is preserved.

Khái niệm cốt lõi

A **Normal distribution** is symmetric, unimodal, and bell-shaped, fully described by its mean μ (the center, where the peak sits) and standard deviation σ (which controls how wide/flat or narrow/tall the bell is). We write $N(\mu, \sigma)$. Many real variables — heights, standardized test scores, measurement errors — are approximately Normal.

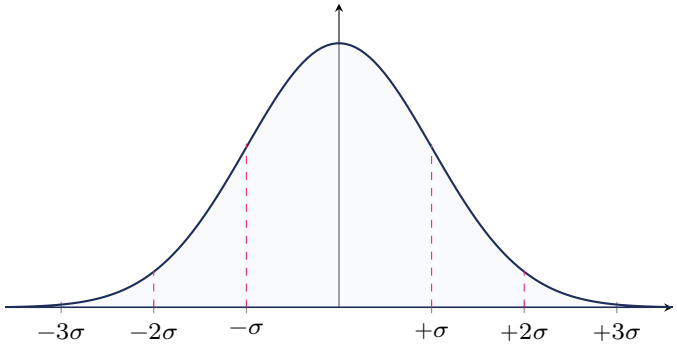
VN **GIẢI THÍCH TIẾNG VIỆT**

Đường cong mật độ (density curve) luôn nằm trên hoặc bằng trục hoành và có tổng diện tích bên dưới bằng 1 (tức 100% dữ liệu); diện tích trên một khoảng chính là tỉ lệ dữ liệu rơi vào khoảng đó. **Phân phối chuẩn (Normal distribution)** có hình chuông, đối xứng, một đỉnh, được xác định hoàn toàn bởi trung bình μ (vị trí đỉnh) và độ lệch chuẩn σ (độ "béo/gầy" của chuông).

For any density curve, the **median** is the point that divides the area exactly in half, while the **mean** is the curve's "balance point" (where it would balance if cut from solid material). For a symmetric curve such as the Normal curve, these two points coincide exactly at the center, which is also where the peak (mode) sits — so for a Normal distribution, mean $=$ median $=$ mode. For a skewed density curve, by contrast, the mean is pulled toward the longer tail while the median is not, exactly mirroring the behavior of skewed data distributions discussed earlier in this chapter.

1.7.2 The Empirical Rule (68–95–99.7)

Empirical Rule (68–95–99.7): for a Normal distribution, about 68% of values fall within 1σ of μ, about 95% within 2σ, and about 99.7% within 3σ.



The Normal curve with the 68–95–99.7 rule marked at $\pm 1\sigma, \pm 2\sigma$.

VN GIẢI THÍCH TIẾNG VIỆT

Quy tắc 68–95–99.7 giúp ước lượng nhanh mà không cần bảng z: khoảng $\mu \pm 1\sigma$ chứa khoảng 68% dữ liệu, $\mu \pm 2\sigma$ chứa khoảng 95%, $\mu \pm 3\sigma$ chứa khoảng 99.7%.

Bài tập luyện tập

Adult male heights are approximately Normal with $\mu = 70$ in and $\sigma = 3$ in. About what percent of men are *taller* than 76 inches?

(A) 0.15%
(B) 2.5%
(C) 5%
(D) 16%

Lời giải chi tiết

76 in is $\dfrac{76-70}{3} = 2$ standard deviations above the mean. By the 68–95–99.7 rule, 95% of men are within $\pm 2\sigma$, leaving 5% split evenly between the two tails: 2.5% above 76 in. **(B)**.

1.7.3 z-Scores and Standardization

A **z-score** standardizes a value: $z = \dfrac{x-\mu}{\sigma}$, giving the number of standard deviations x is from the mean. Standardizing lets us compare values from *different* Normal distributions on a common scale.

VN GIẢI THÍCH TIẾNG VIỆT

Điểm chuẩn hoá $z = \dfrac{x-\mu}{\sigma}$ cho biết x cách trung bình bao nhiêu độ lệch chuẩn — z âm nghĩa là x nhỏ hơn trung bình, z dương nghĩa là x lớn hơn trung bình. Vì z-score không có đơn vị, ta có thể dùng nó để so sánh hai giá trị đến từ hai phân phối chuẩn khác nhau.

(!) Lỗi thường gặp: Độ lệch chuẩn **mẫu** luôn chia cho $n-1$ (không phải n) trong AP Statistics — máy tính CASIO/TI có cả hai (σ_x và s_x); luôn dùng s_x trừ khi đề nói rõ đây là *toàn bộ* population.

1.7.4 Computing Normal Probabilities from a Table

The empirical rule only handles values exactly 1, 2, or 3 standard deviations from the mean. For any other value, we standardize to a z-score and look up $\Phi(z) = P(Z < z)$, the proportion of a standard Normal distribution ($\mu = 0, \sigma = 1$) below z, in a table of the standard Normal cumulative distribution.

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.00	0.5000	1.04	0.8508	1.70	0.9554
0.25	0.5987	1.05	0.8531	1.80	0.9641
0.50	0.6915	1.09	0.8621	1.90	0.9713
0.67	0.7486	1.10	0.8643	1.96	0.9750
0.75	0.7734	1.20	0.8849	2.00	0.9772
0.84	0.7995	1.28	0.8997	2.33	0.9901
1.00	0.8413	1.645	0.9500	3.00	0.9987

Selected values of the standard Normal cumulative distribution $\Phi(z) = P(Z < z)$.

To find the proportion **below** x: standardize, $z = \dfrac{x-\mu}{\sigma}$, then read $\Phi(z)$ directly. To find the proportion **above** x: compute $1-\Phi(z)$. To find the proportion **between** two values $x_1 < x_2$: compute $\Phi(z_2) - \Phi(z_1)$. Because the standard Normal curve is symmetric about 0, $\Phi(-z) = 1-\Phi(z)$.

Empirical rule vs. table: a consistency check. The empirical rule says about 95% of values lie within $\pm 2\sigma$, leaving 5% in the two tails combined. Using the precise table value instead, $\Phi(2.00) = 0.9772$, so the exact proportion in the upper tail beyond 2σ is $1-0.9772 = 0.0228 = 2.28\%$, and $2 \times 2.28\% = 4.56\%$ in both tails combined — close to, but not exactly, the empirical rule's rounded 5%. The empirical rule is a convenient, memorable *approximation* for whole-number multiples of σ; the standard Normal table gives the precise value for any z.

Ví dụ mẫu · Area below a value

Adult women's heights are approximately Normal with $\mu = 64$ in, $\sigma = 2.5$ in. Find the proportion of women shorter than 68 in.

$z = \dfrac{68-64}{2.5} = \dfrac{4}{2.5} = 1.60$. From the table, $\Phi(1.60) = 0.9452$. About 94.52% of women are shorter than 68 inches.

Ví dụ mẫu · Area between two values

Using the same distribution ($\mu = 64$, $\sigma = 2.5$ in), find the proportion of women between 60 and 68 inches tall.

$z_1 = \dfrac{60-64}{2.5} = -1.60 \Rightarrow \Phi(-1.60) = 1 - 0.9452 = 0.0548.$ $\quad z_2 = \dfrac{68-64}{2.5} = 1.60 \Rightarrow \Phi(1.60) = 0.9452.$

$$P(60 < x < 68) = \Phi(1.60) - \Phi(-1.60) = 0.9452 - 0.0548 = 0.8904.$$

About 89.04% of women are between 60 and 68 inches tall.

(!) Lỗi thường gặp: Bảng chuẩn tắc cho $\Phi(z) = P(Z < z)$ — diện tích **bên trái** của z. Muốn tính $P(Z > z)$ (bên phải), phải lấy $1-\Phi(z)$; muốn tính diện tích **giữa** hai giá trị, lấy $\Phi(z_{\text{trên}}) - \Phi(z_{\text{dưới}})$. Quên trừ cho 1 khi đề hỏi "lớn hơn" hoặc "trên" là lỗi rất phổ biến.

Ví dụ mẫu · The IQR of a Normal distribution

Using the standard Normal table, find the interquartile range of the women's-height distribution ($\mu = 64$, $\sigma = 2.5$ in).

Q_1 is the 25th percentile: we need z with $\Phi(z) = 0.25$. By symmetry this is the negative of the z for the 75th percentile; the table gives $\Phi(0.67) = 0.7486 \approx 0.75$, so Q_1 corresponds to $z = -0.67$ and Q_3 corresponds to $z = 0.67$.

$Q_1 = 64 + (-0.67)(2.5) = 64 - 1.675 = 62.33$ in, $\quad Q_3 = 64 + (0.67)(2.5) = 64 + 1.675 = 65.68$ in.

$$IQR = Q_3 - Q_1 = 65.68 - 62.33 \approx 3.35 \text{ in.}$$

In general, for *any* Normal distribution, $IQR \approx 2(0.67)\sigma = 1.34\sigma$ — a useful shortcut linking the Normal model directly back to the IQR studied earlier in this chapter.

Ví dụ mẫu · A symmetric middle-percentage interval

What interval of heights contains the **middle** 80% of adult women, again using $\mu = 64$, $\sigma = 2.5$ in?

"Middle 80%" leaves 20% split evenly between the two tails, i.e. 10% in each tail. We need z with $\Phi(z) = 0.90$ (leaving 10% above); the table gives $\Phi(1.28) = 0.8997 \approx 0.90$, so $z = \pm 1.28$ marks the boundaries.

$$x_{\text{low}} = 64 - 1.28(2.5) = 64 - 3.2 = 60.8 \text{ in}, \qquad x_{\text{high}} = 64 + 1.28(2.5) = 64 + 3.2 = 67.2 \text{ in}.$$

The middle 80% of women have heights between 60.8 and 67.2 inches. (Notice 67.2 in matches the 90th-percentile cutoff found earlier — as it must, since the upper boundary of the middle 80% *is*, by definition, the 90th percentile.)

1.7.5 Inverse Normal: Finding a Value from a Percentile

Sometimes we know a *proportion* or *percentile* and must work backward to find the corresponding value of x — an **inverse Normal** problem. The procedure is reversed: find the z whose table value $\Phi(z)$ matches the target proportion, then solve $z = \dfrac{x - \mu}{\sigma}$ for x:

$$x = \mu + z\sigma.$$

Ví dụ mẫu · Inverse Normal

Using the same women's-height distribution ($\mu = 64$, $\sigma = 2.5$ in), find the height that marks the 90th percentile.

We need z such that $\Phi(z) = 0.90$. From the table, $\Phi(1.28) = 0.8997 \approx 0.90$, so $z \approx 1.28$.

$$x = \mu + z\sigma = 64 + 1.28(2.5) = 64 + 3.2 = 67.2 \text{ in}.$$

A woman would need to be at least about 67.2 inches tall to be in the tallest 10% of women.

1.7.6 Assessing Normality

Before applying Normal-based methods, it is good practice to check whether data are actually approximately Normal.

Normal probability plot

A **Normal probability plot** graphs each data value against its expected z-score if the data were exactly Normal. If the data *are* approximately Normal, the points fall close to a straight line. Systematic **curvature** (the points bending away from a line) indicates the data are skewed; points that veer sharply away from the line only at the two ends indicate outliers or heavier-than-Normal tails.

VN **GIẢI THÍCH TIẾNG VIỆT**

Biểu đồ xác suất chuẩn (Normal probability plot) vẽ từng giá trị dữ liệu so với z-score kỳ vọng nếu dữ liệu phân phối chuẩn hoàn hảo. Nếu các điểm nằm gần một **đường thẳng**, dữ liệu **xấp xỉ chuẩn**. Nếu đồ thị bị **cong** một cách hệ thống, dữ liệu bị lệch (skewed); nếu chỉ hai đầu lệch khỏi đường thẳng, có thể do ngoại lệ hoặc đuôi phân phối "dày" hơn phân phối chuẩn.

Ví dụ mẫu · Reading a Normal probability plot

Two data sets each produce a Normal probability plot. Plot 1's points fall almost exactly on a straight line from corner to corner. Plot 2's points follow a line through the middle but bend noticeably upward at the right end, with the single largest data value sitting well above where the line would predict. Interpret each plot.

Plot 1 indicates the data are **approximately Normal** — no important information is lost by using the empirical rule or table-based Normal calculations on this variable.

Plot 2's upward bend at the right end signals that the largest value(s) sit farther from the center than a Normal model would predict — consistent with either a **right-skewed** distribution or a single high **outlier**. Methods built on the assumption of Normality (such as the empirical rule) should be applied cautiously, if at all, to this data set; a five-number summary and boxplot would likely describe it more faithfully.

1.8 Practice problems

Bài tập luyện tập

A survey of 50 cars parked in a lot records their color: Red 15, Blue 20, White 10, Black 5. What percent of the cars are Blue?

(A) 20%
(C) 40%
(B) 30%
(D) 50%

Lời giải chi tiết

Relative frequency of Blue $= \dfrac{20}{50} = 0.40 = 40\%$. **(C)**.

Bài tập luyện tập

Which type of graph is best suited to precisely compare the counts of six categories against one another (rather than emphasizing each category's share of a single whole)?

(A) Pie chart
(C) Boxplot
(B) Bar graph
(D) Normal probability plot

Lời giải chi tiết

A **bar graph** places every category's count on the same numeric axis, so heights can be compared directly and precisely; a pie chart emphasizes parts of a whole and is harder to compare precisely once there are several categories. **(B)**.

Bài tập luyện tập

A dotplot of number of siblings for 12 students shows: 0 siblings (3 students), 1 sibling (4 students), 2 siblings (3 students), 3 siblings (1 student), 4 siblings (1 student). What is the mode of this distribution?

(A) 0
(B) 1
(C) 2
(D) 3

Lời giải chi tiết

The mode is the most frequent value. Here 1 sibling occurs for 4 students, more than any other value ($3, 4, 3, 1, 1$ students respectively for $0, 1, 2, 3, 4$ siblings). **(B)**.

Bài tập luyện tập

The stemplot below shows the ages of 12 employees at a small company.

Stem	Leaves
2	3 5 8
3	1 2 4 6 9
4	0 3 5
5	2

Find the median and the IQR of the ages, and determine whether age 52 is a suspected outlier.

Lời giải chi tiết

Reading off the stemplot, the sorted ages are $23, 25, 28, 31, 32, 34, 36, 39, 40, 43, 45, 52$ ($n = 12$). Median = average of 6th and 7th values $= \dfrac{34+36}{2} = 35$. Lower half $\{23, 25, 28, 31, 32, 34\} \Rightarrow Q_1 = \dfrac{28+31}{2} = 29.5$. Upper half $\{36, 39, 40, 43, 45, 52\} \Rightarrow Q_3 = \dfrac{40+43}{2} = 41.5$. $IQR = 41.5 - 29.5 = 12$. Upper fence $= 41.5 + 1.5(12) = 59.5$. Since $52 < 59.5$, age 52 is **not** a suspected outlier.

Bài tập luyện tập

Two commuting groups are summarized by five-number summaries (minutes): Bus riders — min 10, $Q_1 = 18$, median 25, $Q_3 = 32$, max 55. Car drivers — min 8, $Q_1 = 12$, median 15, $Q_3 = 20$, max 40. Compare the two distributions using SOCS, checking each for outliers with the $1.5 \times IQR$ rule.

Lời giải chi tiết

Bus riders: $IQR = 32 - 18 = 14$; upper fence $= 32 + 1.5(14) = 53$; since $55 > 53$, the maximum is an **outlier**. Lower fence $= 18 - 21 = -3$; the minimum (10) is not below it.
Car drivers: $IQR = 20 - 12 = 8$; upper fence $= 20 + 1.5(8) = 32$; since $40 > 32$, the maximum is an **outlier**. Lower fence $= 12 - 12 = 0$; the minimum (8) is not below it.
Comparison: the median commute is much shorter for car drivers (15 min) than bus riders (25 min), and bus commute times are more variable ($IQR = 14$ min vs. 8 min for cars). Both groups have an occasional unusually long commute (a high outlier), but otherwise cars are both faster and more consistent.

Bài tập luyện tập

In a back-to-back stemplot comparing two classes' quiz scores, Class 1's leaves are spread across many stems while Class 2's leaves cluster tightly around a single stem. Which conclu-

sion is correct?

(A) Class 1 has smaller spread than Class 2
(B) Class 2 has smaller spread than Class 1
(C) Both classes have equal spread
(D) The stemplot cannot show spread

Lời giải chi tiết

Leaves spread across many stems indicate values scattered over a wide range (larger spread); leaves clustered around one stem indicate values packed closely together (smaller spread). **(B).**

Bài tập luyện tập

A histogram of number of pets owned by families shows most families owning 0–2 pets, with a long tail extending toward families owning 6–10 pets. This shape is best described as:

(A) Skewed left
(B) Skewed right
(C) Symmetric
(D) Uniform

Lời giải chi tiết

The long tail points toward the larger values (6–10 pets), which is the defining feature of a **skewed right** distribution. **(B).**

Bài tập luyện tập

A distribution of birth months among 120 people shows roughly the same frequency (about 10 people) in every one of the 12 months, with no distinct peak. This shape is:

(A) Skewed right
(B) Bimodal
(C) Uniform
(D) Unimodal and symmetric

Lời giải chi tiết

Roughly equal frequency across the entire range, with no clear peak, is the definition of a **uniform** distribution. **(C).**

Bài tập luyện tập

Data set A: $\{40, 42, 43, 45, 120\}$. Which measure of center best describes a typical value, and why?

(A) Mean, because it uses every value
(B) Median, because 120 is a resistant outlier
(C) Median, because 120 is a strong outlier and the median is resistant
(D) Standard deviation, because it measures spread

Lời giải chi tiết

120 is far from the rest of the data (an outlier), which would pull the mean upward and make it unrepresentative. The **median** is resistant to outliers, so it better describes a typical value. **(C).**

Bài tập luyện tập

Seven donors give the following amounts ($): $10, 15, 15, 20, 20, 25, 200$. Compute the mean and the median, and describe the shape of the distribution and which measure of center is more appropriate.

Lời giải chi tiết

Sum $= 10 + 15 + 15 + 20 + 20 + 25 + 200 = 305$; mean $= \dfrac{305}{7} \approx 43.57$. Sorted data is already given; median (4th of 7 values) $= 20$. Since mean (≈ 43.57) is much greater than the median (20), the distribution is strongly **skewed right**, driven by the $200 outlier. The **median** ($20) better represents a typical donation; the mean is inflated by the single large gift.

Bài tập luyện tập

If the largest value in a data set is increased (and remains the largest value), which of the following statistics *must* increase?

(A) The median
(B) The mean
(C) The mode
(D) The IQR

Lời giải chi tiết

The mean directly incorporates the exact size of every value, so increasing any value (while n stays fixed) strictly increases the sum and therefore the **mean**. The median, mode, and IQR depend only on the *position* (rank) of values, not their exact size, and since the value stays the largest, none of these positional statistics change. **(B)**.

Bài tập luyện tập

Exam scores for 5 students are $70, 75, 80, 85, 90$. Compute the sample variance and the sample standard deviation.

Lời giải chi tiết

Mean $= \dfrac{70+75+80+85+90}{5} = \dfrac{400}{5} = 80$. Deviations: $-10, -5, 0, 5, 10$; squared: $100, 25, 0, 25, 100$, summing to 250.

$$s_x^2 = \frac{250}{5-1} = \frac{250}{4} = 62.5, \qquad s_x = \sqrt{62.5} \approx 7.91.$$

Bài tập luyện tập

Verify $Q_1 = 7$ and $Q_3 = 11.5$ for the data set $\{5, 7, 7, 8, 9, 10, 13, 15\}$ using the exclude-the-median method. Is the outlier rule triggered by 15? Show all fence calculations.

Lời giải chi tiết

$n = 8$ (even), so split evenly: lower half $\{5, 7, 7, 8\} \Rightarrow Q_1 = \frac{7+7}{2} = 7$; upper half $\{9, 10, 13, 15\} \Rightarrow Q_3 = \frac{10+13}{2} = 11.5$. So $IQR = 11.5 - 7 = 4.5$. Upper fence $= 11.5 + 1.5(4.5) = 11.5 + 6.75 = 18.25$. Since $15 < 18.25$, 15 **is NOT a suspected outlier**.

Lower fence $= 7 - 6.75 = 0.25$; no values fall below it either.

Bài tập luyện tập

Find the five-number summary of the data set $\{12, 15, 18, 18, 20, 22, 25, 50\}$, check for outliers, and describe how the modified boxplot's upper whisker would be drawn.

Lời giải chi tiết

$n = 8$ (even). Median $= \dfrac{18+20}{2} = 19$. Lower half $\{12, 15, 18, 18\} \Rightarrow Q_1 = \dfrac{15+18}{2} = 16.5$. Upper half $\{20, 22, 25, 50\} \Rightarrow Q_3 = \dfrac{22+25}{2} = 23.5$. $IQR = 23.5 - 16.5 = 7$. Upper fence $= 23.5 + 1.5(7) = 34$; since $50 > 34$, it is a **suspected outlier**. Lower fence $= 16.5 - 1.5(7) = 6$; the minimum (12) is not below it. Because 50 is an outlier, the upper whisker stops at the largest *non-outlier* value, 25, and 50 is plotted as a separate point.

Bài tập luyện tập

A data set has $Q_1 = 40$ and $Q_3 = 55$. Which of the following values would be flagged as a *low* outlier by the $1.5 \times IQR$ rule?

(A) 20
(B) 17
(C) 42
(D) 55

Lời giải chi tiết

$IQR = 55 - 40 = 15$. Lower fence $= 40 - 1.5(15) = 40 - 22.5 = 17.5$. Only 17 is below 17.5 (note $20 > 17.5$, so it is not flagged). **(B)**.

Bài tập luyện tập

A set of temperatures has mean $68\,^\circ$F and standard deviation $5\,^\circ$F. If every temperature is converted by subtracting a constant offset of $10\,^\circ$F, what are the new mean and standard deviation?

(A) Mean $58\,^\circ$F, SD $5\,^\circ$F
(B) Mean $58\,^\circ$F, SD $-5\,^\circ$F
(C) Mean $68\,^\circ$F, SD $5\,^\circ$F
(D) Mean $58\,^\circ$F, SD $15\,^\circ$F

Lời giải chi tiết

Subtracting a constant is adding $a = -10$: the mean shifts to $68 - 10 = 58\,^\circ$F, but the standard deviation is **unaffected** by a shift, so it remains $5\,^\circ$F. **(A)**.

Bài tập luyện tập

A set of distances (miles) has mean 12, median 10, range 20, and standard deviation 4. Convert every distance to kilometers by multiplying by 1.6. Find the new mean, median, range, and standard deviation (in km).

Lời giải chi tiết

Multiplying by a positive constant scales every one of these statistics by that same constant: new mean $= 12(1.6) = 19.2$ km; new median $= 10(1.6) = 16$ km; new range $= 20(1.6) = 32$ km; new standard deviation $= 4(1.6) = 6.4$ km.

Bài tập luyện tập

Pumpkin weights are approximately Normal with mean 8 lb and standard deviation 1.5 lb. What percent of pumpkins weigh between 5 lb and 11 lb?

(A) 68%
(B) 95%
(C) 99.7%
(D) 50%

Lời giải chi tiết

$5 = 8 - 2(1.5)$ and $11 = 8 + 2(1.5)$, so this interval is exactly $\mu \pm 2\sigma$. By the empirical rule, this contains about 95% of pumpkins. **(B)**.

Bài tập luyện tập

A student scored 82 on a test with mean 75 and standard deviation 5. Find the student's z-score and the approximate percentile this corresponds to.

(A) $z = 1.4$, ≈ 92nd percentile
(B) $z = 1.4$, ≈ 84th percentile
(C) $z = 0.7$, ≈ 92nd percentile
(D) $z = 1.4$, ≈ 50th percentile

Lời giải chi tiết

$z = \dfrac{82 - 75}{5} = \dfrac{7}{5} = 1.4$. From the standard Normal table, $\Phi(1.4) = 0.9192$, i.e. about the 92nd percentile. **(A)**.

Bài tập luyện tập

Battery lifetimes are approximately Normal with mean 500 hours and standard deviation 25 hours. Find the proportion of batteries lasting less than 550 hours.

Lời giải chi tiết

$z = \dfrac{550 - 500}{25} = \dfrac{50}{25} = 2.00$. From the table, $\Phi(2.00) = 0.9772$. About 97.72% of batteries last less than 550 hours.

Bài tập luyện tập

Using the same battery distribution (mean 500 hours, SD 25 hours), find the proportion of batteries lasting between 460 and 540 hours.

Lời giải chi tiết

$z_1 = \dfrac{460-500}{25} = -1.60 \Rightarrow \Phi(-1.60) = 1 - 0.9452 = 0.0548.$ $z_2 = \dfrac{540-500}{25} = 1.60 \Rightarrow \Phi(1.60) = 0.9452.$

$$P(460 < x < 540) = 0.9452 - 0.0548 = 0.8904 \approx 89.0\%.$$

Bài tập luyện tập

IQ scores are approximately Normal with mean 100 and standard deviation 15. What IQ score marks the 84th percentile?

(A) 100
(B) 108
(C) 115
(D) 130

Lời giải chi tiết

We need z with $\Phi(z) = 0.84$; the table gives $\Phi(1.00) = 0.8413 \approx 0.84$, so $z \approx 1.00$. $x = 100 + 1.00(15) = 115$. **(C)**.

Bài tập luyện tập

Cholesterol levels are approximately Normal with mean 190 mg/dL and standard deviation 20 mg/dL. A clinic wants to flag the lowest 5% of patients for follow-up. What cholesterol level marks this cutoff?

Lời giải chi tiết

The lowest 5% corresponds to $\Phi(z) = 0.05$. By symmetry, this is the negative of the z for the upper 95%: since $\Phi(1.645) = 0.9500$, we have $z = -1.645$.

$$x = 190 + (-1.645)(20) = 190 - 32.9 = 157.1 \text{ mg/dL}.$$

Patients with cholesterol below approximately 157.1 mg/dL would be flagged.

Bài tập luyện tập

Package weights are approximately Normal with mean 50 g and standard deviation 4 g. What weight marks the 99th percentile?

(A) 54 g
(B) 58 g
(C) 59.3 g
(D) 66 g

Lời giải chi tiết

We need z with $\Phi(z) = 0.99$; the table gives $\Phi(2.33) = 0.9901 \approx 0.99$, so $z \approx 2.33$. $x = 50 + 2.33(4) = 50 + 9.32 = 59.32 \approx 59.3$ g. **(C)**.

Bài tập luyện tập

A Normal probability plot of a data set curves noticeably and is not close to a straight line. What does this indicate about the data?

(A) The data are approximately Normal
(B) The data are not approximately Normal (a departure such as skewness)
(C) The data have no outliers
(D) The data must be categorical

Lời giải chi tiết

A straight-line pattern on a Normal probability plot indicates approximate Normality; systematic curvature is exactly the signal that the data **depart** from a Normal shape (e.g., skewness). **(B)**.

Bài tập luyện tập

On Test A (mean 70, SD 8), Maria scored 82. On Test B (mean 100, SD 15), Julia scored 120. Who performed better relative to her own class?

(A) Maria, since her z-score (1.5) is higher than Julia's (≈ 1.33)
(B) Julia, since her raw score is higher
(C) They performed equally well
(D) Cannot be determined without more information

Lời giải chi tiết

$z_{\text{Maria}} = \dfrac{82-70}{8} = 1.5$. $z_{\text{Julia}} = \dfrac{120-100}{15} = 1.333$. Maria's score is farther above her class's mean in standard-deviation units, so she performed relatively better, even though Julia's raw score is higher. **(A)**.

Bài tập luyện tập

A data set has mean 42 and median 37. The distribution is most likely:

(A) Skewed left
(B) Skewed right
(C) Perfectly symmetric
(D) Cannot be determined

Lời giải chi tiết

Mean $>$ median means the mean has been pulled upward by a long right tail, so the distribution is **skewed right**. **(B)**.

Bài tập luyện tập

A city records resident complaints this month: Noise $= 45$, Parking $= 30$, Trash $= 15$, Other $= 10$. Which type of display orders these categories from most to least frequent, and what is that display called?

(A) A pie chart, called a dot chart
(B) A bar graph with categories sorted by frequency, called a Pareto chart
(C) A histogram, called a frequency polygon
(D) A boxplot, called an ordinal chart

Lời giải chi tiết

Ordering bars from tallest (most frequent) to shortest (least frequent) is exactly the definition of a **Pareto chart** — a bar graph, not a pie chart, histogram, or boxplot, with bars sorted by descending frequency. **(B)**.

Bài tập luyện tập

For the data set $\{5, 8, 9, 10, 11, 12, 14, 40\}$, find Q_1, Q_3, and IQR, then determine whether 40 is not an outlier, a mild outlier, or an extreme outlier.

Lời giải chi tiết

$n = 8$ (even). Lower half $\{5, 8, 9, 10\} \Rightarrow Q_1 = \dfrac{8+9}{2} = 8.5$. Upper half $\{11, 12, 14, 40\} \Rightarrow Q_3 = \dfrac{12+14}{2} = 13$. $IQR = 13 - 8.5 = 4.5$. Mild upper fence $= 13 + 1.5(4.5) = 19.75$; since $40 > 19.75$, 40 is at least a mild outlier. Extreme upper fence $= 13 + 3(4.5) = 26.5$; since $40 > 26.5$ as well, 40 is an **extreme outlier**.

Exploring Two-Variable Data

Mục tiêu Unit: Bảng hai chiều (two-way table), phân bố biên (marginal distribution) và phân bố có điều kiện (conditional distribution) cho dữ liệu định tính hai biến, biểu đồ cột phân đoạn (segmented bar chart); biểu đồ phân tán (scatterplot) và hệ số tương quan r; đường hồi quy bình phương tối thiểu (LSRL) — hệ số góc, hệ số chặn, dự đoán và nguy cơ ngoại suy (extrapolation); phần dư (residual) và biểu đồ phần dư để kiểm tra tính tuyến tính; hệ số xác định r^2 và độ lệch chuẩn phần dư s; phân biệt điểm ngoại lệ (outlier) và điểm ảnh hưởng (influential point) trong hồi quy; biến đổi dữ liệu bằng logarit (log, log-log) để tuyến tính hoá mô hình mũ và mô hình lũy thừa.

2.1 Bivariate Categorical Data: Two-Way Tables

Core idea

Two-variable (bivariate) data comes in two flavors, and the tool you use to explore the relationship depends on the type of each variable. When **both variables are categorical**, relationships are explored with **two-way tables** and **conditional distributions**. When **both variables are quantitative**, relationships are explored with **scatterplots**, **correlation**, and **regression** – the tools that occupy the rest of this unit.

VN **GIẢI THÍCH TIẾNG VIỆT**

Dữ liệu hai biến có hai dạng: nếu cả hai biến đều là **định tính (categorical)** (VD giới tính, phương tiện di chuyển), ta dùng **bảng hai chiều** và **phân bố có điều kiện** để tìm mối liên hệ. Nếu cả hai biến đều là **định lượng (quantitative)** (VD chiều cao, điểm số), ta dùng **biểu đồ phân tán**, **hệ số tương quan**, và **hồi quy tuyến tính** – nội dung chiếm phần lớn chương này.

2.1.1 Two-way tables and marginal distributions

A **two-way table** (or contingency table) displays counts of individuals classified by two categorical variables simultaneously – one variable defines the rows, the other defines the columns. Each entry in the body of the table is a **joint count**: the number of individuals falling into that specific combination of row category and column category. Row and column totals (and the grand total in the corner) are added around the edges of the table.

Two-way table: class year and transportation

A college surveys 400 students, recording each student's class year (Underclassman or Upperclassman) and primary mode of transportation to campus (Walk, Drive, or Bus).

	Walk	Drive	Bus	Total
Underclassmen	90	60	50	200
Upperclassmen	30	140	30	200
Total	120	200	80	400

The **marginal distribution** of a single variable is found from the totals in the margins of the table (as percentages of the grand total), ignoring the other variable entirely.

Marginal distribution of class year: Underclassmen $= \frac{200}{400} = 50\%$, Upperclassmen $= \frac{200}{400} = 50\%$.

Marginal distribution of transportation: Walk $= \frac{120}{400} = 30\%$, Drive $= \frac{200}{400} = 50\%$, Bus $= \frac{80}{400} = 20\%$.

Marginal distribution: the distribution of one categorical variable alone, computed from the row totals or column totals divided by the *grand total*. It says nothing about the relationship between the two variables. **Joint distribution:** the distribution of the combination of both variables at once – each cell count divided by the grand total.

VN **GIẢI THÍCH TIẾNG VIỆT**

Phân bố biên (marginal distribution) là phân bố của MỘT biến duy nhất, tính từ tổng hàng hoặc tổng cột chia cho tổng lớn (grand total) – hoàn toàn bỏ qua biến còn lại. **Phân bố kết hợp (joint distribution)** là phân bố của cặp giá trị (biến 1, biến 2) đồng thời – mỗi ô trong bảng chia cho tổng lớn. Học sinh thường nhầm hai khái niệm này với **phân bố có điều kiện** (xem mục tiếp theo) – ba loại phân bố này có mẫu số khác nhau: phân bố biên chia cho tổng lớn, phân bố kết hợp cũng chia cho tổng lớn (nhưng ở từng ô), còn phân bố có điều kiện chia cho MỘT tổng hàng hoặc MỘT tổng cột cụ thể.

2.1.2 Conditional distributions and association

A **conditional distribution** of one variable is the distribution of that variable restricted to just the individuals who fall into a single category of the other variable – each relevant cell count is divided by that category's row (or column) total, *not* the grand total.

To find the **conditional distribution** of variable Y given a specific category of variable X: take each joint count in that row (or column) and divide by the row's (or column's) total. Two categorical variables show an **association** if the conditional distributions of Y differ noticeably across the categories of X. If the conditional distributions are nearly identical, the variables show **no association** (they behave independently).

Two categorical variables are said to be **independent** if knowing the category of one variable gives no additional information about the other – equivalently, every conditional distribution of Y (across all categories of X) matches the overall marginal distribution of Y. Independence is therefore the categorical-data counterpart of "no association": the opposite of independence is exactly the kind of association this section is built to detect.

Conditional distributions and interpreting association

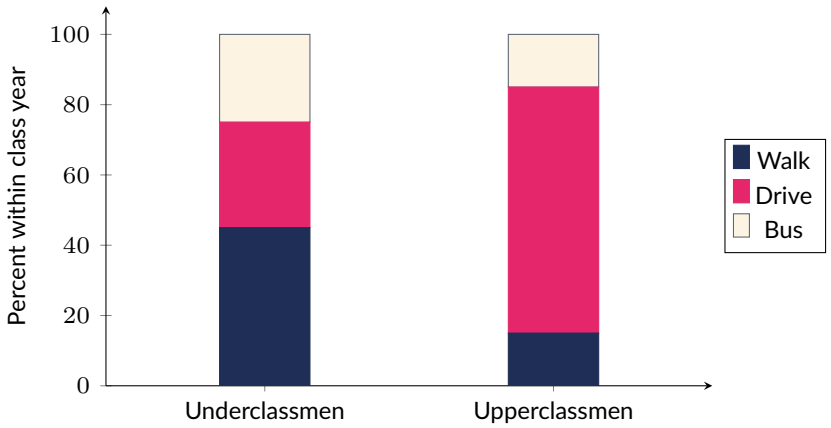
Using the class-year/transportation table above, find the conditional distribution of transportation mode given Underclassman, and given Upperclassman, and describe the association.

Given Underclassman (divide the top row by its row total, 200): Walk $= \frac{90}{200} = 45\%$, Drive $= \frac{60}{200} = 30\%$, Bus $= \frac{50}{200} = 25\%$.

Given Upperclassman (divide the bottom row by its row total, 200): Walk $= \frac{30}{200} = 15\%$, Drive

$= \frac{140}{200} = 70\%$, Bus $= \frac{30}{200} = 15\%$.

The two conditional distributions ($45/30/25$ vs. $15/70/15$) are very different – Underclassmen walk about three times as often as Upperclassmen (45% vs. 15%), while Upperclassmen drive more than twice as often (70% vs. 30%). This is strong evidence of an **association** between class year and transportation mode: knowing a student's class year changes what you would predict about how they get to campus.



Segmented bar chart of the conditional distribution of transportation mode, given class year. The very different segment heights across the two bars visually confirm the association between class year and transportation mode.

(!) Lỗi thường gặp: Đừng chia cho sai tổng. Khi tính **phân bố có điều kiện** của Y cho một hàng cụ thể, phải chia cho **tổng của hàng đó** (VD 200), không phải tổng lớn (400). Ngược lại, khi tính **phân bố biên**, phải chia tổng hàng/cột cho **tổng lớn**. Nhầm mẫu số là lỗi phổ biến nhất khi làm bài về bảng hai chiều. Ngoài ra, để kết luận có **association** hay không, phải so sánh các **phân bố có điều kiện** với nhau (không phải so sánh các con số tuyệt đối, vì các nhóm có thể có cỡ mẫu khác nhau).

Conditioning in the other direction

Using the same class-year/transportation table, condition on transportation mode instead of class year: find the distribution of class year given Walk, given Drive, and given Bus.

Given Walk (divide the Walk column by its total, 120): Underclassmen $= \frac{90}{120} = 75\%$, Upperclassmen $= \frac{30}{120} = 25\%$.

Given Drive (divide the Drive column by its total, 200): Underclassmen $= \frac{60}{200} = 30\%$, Upperclassmen $= \frac{140}{200} = 70\%$.

Given Bus (divide the Bus column by its total, 80): Underclassmen $= \frac{50}{80} = 62.5\%$, Upperclassmen $= \frac{30}{80} = 37.5\%$.

Walkers are overwhelmingly Underclassmen (75%), while drivers are overwhelmingly Upperclassmen (70%) – the same underlying association shows up regardless of which variable is used to condition, though the specific percentages differ depending on the direction chosen.

Three distributions, three denominators. It helps to keep the three distributions that can be read from a two-way table straight, since each divides by a different total.

Distribution	Denominator used	What it describes
Marginal	the grand total	the distribution of ONE variable alone (row totals or column totals over the grand total)
Joint	the grand total	the distribution of a specific (row category, column category) *pair* – one cell over the grand total
Conditional	one row total or one column total	the distribution of one variable, restricted to individuals in a single category of the other variable

Row percentages and column percentages are simply conditional distributions computed in the two possible directions – conditioning on the row variable (dividing across each row by its row total) or conditioning on the column variable (dividing down each column by its column total). Either direction is valid; the appropriate choice depends on which variable is naturally thought of as explaining the other.

Association between two categorical variables can also be displayed graphically. A **segmented bar chart** (as above) stacks each category's conditional distribution into a single bar reaching 100%, making it easy to compare shares directly. A **side-by-side (clustered) bar chart** instead draws a separate bar for each category, placed next to each other in groups – useful for comparing raw counts (or percentages) across groups without forcing everything into a single 100% bar. Both displays are built from the same conditional distributions; they simply arrange the same information differently.

As with quantitative variables, an observed association between two categorical variables does not by itself establish that one variable causes the other – a lurking variable could still be responsible (for example, a study might find an association between a student's chosen elective and their exam score, when a lurking variable such as prior interest in the subject actually drives both).

2.2 Scatterplots and Correlation

2.2.1 Describing a scatterplot

A **scatterplot** displays the relationship between two *quantitative* variables measured on the same individuals, with each point representing one individual. We describe a scatterplot using four features:

- **Form**: does the pattern look linear (following a straight line) or curved (following some other shape, e.g. a parabola or exponential curve)?
- **Direction**: does y tend to increase as x increases (**positive** association) or decrease as x increases (**negative** association)?
- **Strength**: how closely do the points cluster around the form? A strong relationship has points tightly hugging a curve or line; a weak relationship has points loosely scattered.
- **Outliers**: any individual points that fall far from the overall pattern, in either the x-direction, the y-direction, or both.

Reasoning: describing a scatterplot in words

A scatterplot plots weekly training mileage (x) against marathon finishing time in minutes (y) for 20 runners. For 19 of the runners, the points decrease from upper left to lower right, hugging a straight line closely. One runner, however, trained very little (low x) yet still finished with a fast time (low y), sitting well below where the pattern would predict. Describe this scatterplot using form, direction, strength, and outliers.

Form: linear (for the main cluster of 19 runners). **Direction:** negative – higher weekly mileage is associated with a faster (lower) finishing time. **Strength:** strong, since the 19 typical points hug the line closely. **Outlier:** the low-mileage runner with an unexpectedly fast time; this single runner does not fit the overall pattern and would pull the LSRL and r away from what the other 19 points alone would suggest.

2.2.2 Explanatory and response variables

When investigating whether one variable might help explain or predict changes in another, the **explanatory variable** (independent variable, x) is the one thought to explain or cause changes, and the **response variable** (dependent variable, y) is the one thought to be affected. By convention, the explanatory variable is plotted on the horizontal axis and the response variable on the vertical axis. Not every pair of variables has a clear explanatory/response role – when there is no such distinction (e.g. height and arm span), either variable may be placed on either axis.

Identifying explanatory and response variables

For each pair, identify a natural explanatory and response variable, if one exists. (a) Daily hours of sunlight and daily ice cream sales at a beach shop. (b) Shoe size and height for a sample of adults.

(a) Hours of sunlight is the natural **explanatory** variable (x) and ice cream sales the **response** (y): sunnier days plausibly drive more ice cream purchases; there is no plausible mechanism running the other direction.
(b) Neither variable clearly explains the other – both shoe size and height are largely determined by a person's overall body size (a common underlying factor). There is no natural causal direction here, so either variable could reasonably be placed on the x-axis.

2.2.3 The correlation coefficient r

The **correlation coefficient** r measures the strength and direction of a *linear* relationship: $-1 \leq r \leq 1$. r is unitless, unaffected by which variable is x or y, and unaffected by unit changes – but r is **not resistant** to outliers and only measures *linear* association.

VN **GIẢI THÍCH TIẾNG VIỆT**
r chỉ đo mức độ **tuyến tính**: r gần ± 1 là quan hệ tuyến tính mạnh, r gần 0 có thể vẫn là quan hệ rất mạnh nhưng **phi tuyến** (ví dụ hình parabol). r không đổi khi đổi đơn vị đo, không có đơn vị, và bị ảnh hưởng mạnh bởi outlier – không "resistant".

Reasoning: estimating r from a description

A scatterplot of a used car's age (in years, x) versus its resale value (in dollars, y) shows points that fall very close to a downward-sloping line, tightly clustered with no obvious curvature and no outliers. Estimate the sign and approximate size of r.

As the car's age increases, resale value decreases – a **negative** association, so $r < 0$. Because the points are described as tightly clustered around a straight line (a strong, clean linear pattern with no outliers pulling the fit off), r should be close to -1 – a reasonable estimate is something like $r \approx -0.90$ to -0.97. (If the description instead said the points were loosely scattered around a downward trend, a more modest magnitude such as $r \approx -0.4$ would be more appropriate.)

How one outlier can transform r

Five points follow a perfectly straight line: $(1, 40)$, $(2, 50)$, $(3, 60)$, $(4, 70)$, $(5, 80)$ (indeed $y = 30 + 10x$ exactly for each point), so $r = 1$ for this data set. A sixth point, $(6, 20)$, is now added – far below where the pattern would predict (the line would predict $y = 30 + 10(6) = 90$ at $x = 6$, but the actual value is only 20).

Recomputing r for all six points gives $r = 0$ exactly: a single poorly-placed point took the correlation from **perfect positive** ($r = 1$) all the way down to **no linear association at all** ($r = 0$). This demonstrates concretely why r is described as **not resistant**: because r is built from every point's squared and cross-product deviations, one point far from the pattern can dominate the calculation and swing r dramatically, especially in a small data set.

2.2.4 Computing r from raw data

The properties above describe how r *behaves*, but r can also be computed directly from raw data using standardized values (z-scores).

$r = \dfrac{1}{n-1}\sum\left(\dfrac{x_i - \bar{x}}{s_x}\right)\left(\dfrac{y_i - \bar{y}}{s_y}\right)$. In words: standardize every x-value and every y-value (convert to z-scores), multiply each pair of z-scores together, sum the products, and divide by $n-1$. Points where x and y are simultaneously above average (or simultaneously below average) contribute a *positive* product; points where one is above average while the other is below contribute a *negative* product – which is why a consistent positive pattern drives r toward $+1$ and a consistent negative pattern drives r toward -1.

Computing r by hand with z-scores

Find r for $(1, 2), (2, 3), (3, 5), (4, 4), (5, 6)$.

$\bar{x} = 3$, $\bar{y} = 4$, and both standard deviations equal $s_x = s_y = \sqrt{2.5} \approx 1.581$ (each set of deviations is $-2, -1, 0, 1, 2$, so both variances equal $\frac{4+1+0+1+4}{4} = 2.5$).

x	y	$z_x = \frac{x-\bar{x}}{s_x}$	$z_y = \frac{y-\bar{y}}{s_y}$	$z_x z_y$
1	2	-1.265	-1.265	1.600
2	3	-0.632	-0.632	0.400
3	5	0.000	0.632	0.000
4	4	0.632	0.000	0.000
5	6	1.265	1.265	1.600

Summing the last column: $1.600+0.400+0.000+0.000+1.600 = 3.600$. So $r = \dfrac{3.600}{5-1} = \dfrac{3.600}{4} = 0.900$ – matching what the strong, tightly-clustered upward pattern in the data suggests.

In practice, r is almost always obtained from a calculator or software rather than by hand, since the z-score computation grows tedious for larger data sets – but the z-score formula is what determines the properties of r discussed earlier (unaffected by units, symmetric in x and y, sensitive to outliers).

GIẢI THÍCH TIẾNG VIỆT (VN)

Công thức $r = \frac{1}{n-1}\sum z_x z_y$ cho thấy bản chất của r: mỗi điểm dữ liệu được "chuẩn hoá" thành hai z-score (đo độ lệch so với trung bình theo đơn vị độ lệch chuẩn), rồi nhân hai z-score lại. Nếu x và y CÙNG cao hơn trung bình (hoặc CÙNG thấp hơn trung bình) tại nhiều điểm, tích

$z_x z_y$ đa số dương $\Rightarrow r$ dương. Nếu một biến cao trong khi biến kia thấp, tích âm $\Rightarrow r$ âm. Đây cũng là lý do r không có đơn vị: z-score đã "khử" đơn vị gốc của x và y.

2.2.5 Correlation, causation, and lurking variables

A strong correlation (or a high r^2) never by itself proves that x causes y. There are three broad explanations for an observed association between x and y: (1) x genuinely causes y; (2) the causation actually runs the other direction, y causes x; or (3) a **lurking (confounding) variable** – a variable not measured in the study – influences both x and y, creating an association between them even though neither causes the other directly.

Reasoning: identifying a lurking variable

Across many cities, the number of firefighters dispatched to a fire and the dollar amount of property damage caused by that fire are strongly positively correlated. Does sending more firefighters cause more damage? Identify a lurking variable that better explains the association.

It is not that firefighters cause damage – rather, the **size (severity) of the fire** is a lurking variable that drives both quantities: a larger, more severe fire requires dispatching more firefighters *and* independently causes more property damage before it can be contained. The size of the fire, not the number of firefighters, is the common cause behind the strong positive correlation.

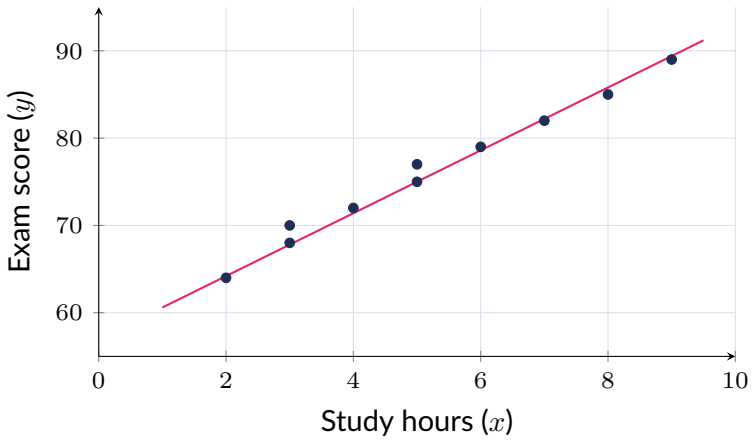
(!) Lỗi thường gặp: Tương quan mạnh (hoặc r^2 cao) **không chứng minh nhân quả (correlation is not causation)**. Hai biến có thể tương quan mạnh vì (1) x thực sự gây ra y, (2) y gây ra x (chiều ngược lại), hoặc (3) một **biến gây nhiễu (lurking/confounding variable)** tác động lên cả hai. Ví dụ kinh điển: số lượng lính cứu hoả được điều đến và thiệt hại của đám cháy tương quan dương mạnh – không phải vì lính cứu hoả gây ra thiệt hại, mà vì **độ lớn của đám cháy** là biến gây nhiễu tác động lên cả hai. Ngoài ra, luôn kiểm tra biểu đồ phần dư (residual plot, xem phần sau): nếu phần dư có **dạng cong**, mô hình tuyến tính không phù hợp dù r có thể vẫn lớn.

2.3 Least-Squares Regression

2.3.1 The regression equation

When a scatterplot shows a reasonably linear pattern, we can summarize it with a line that best fits the data: the **least-squares regression line (LSRL)**.

LSRL: $\hat{y} = a + bx$, where $b = r \cdot \dfrac{s_y}{s_x}$ (slope) and $a = \bar{y} - b\bar{x}$ (intercept). **Residual** = observed − predicted $= y - \hat{y}$. r^2 (coefficient of determination) = the fraction of variation in y explained by the linear model on x. The LSRL is called "least-squares" because, among all possible lines, it is the one that minimizes the sum of the squared residuals, $\sum(y - \hat{y})^2$.



Scatterplot of exam score vs. study hours with the least-squares regression line overlaid: a strong, positive, linear association.

Ví dụ mẫu · Worked example

Summary statistics for study hours (x) and exam score (y): $\bar{x} = 5$, $s_x = 2$, $\bar{y} = 75$, $s_y = 8$, $r = 0.9$.

Slope: $b = 0.9 \cdot \dfrac{8}{2} = 3.6$. Intercept: $a = 75 - 3.6(5) = 75 - 18 = 57$. So $\hat{y} = 57 + 3.6x$. *Interpretation of slope:* for each additional hour studied, predicted exam score increases by about 3.6 points, on average. Predicted score at $x = 6$: $\hat{y} = 57 + 3.6(6) = 78.6$. If the actual score at $x = 6$ was 80, the residual is $80 - 78.6 = 1.4$ (the LSRL *underpredicted*). Also $r^2 = 0.9^2 = 0.81$: about 81% of the variability in exam score is explained by the linear model on study hours.

VN **GIẢI THÍCH TIẾNG VIỆT**

LSRL viết dưới dạng $\hat{y} = a + bx$: b (hệ số góc) cho biết y dự đoán thay đổi bao nhiêu khi x tăng 1 đơn vị; a (hệ số chặn) là giá trị $\hat{y}$ dự đoán khi $x = 0$. Công thức $b = r \cdot s_y/s_x$ nghĩa là hệ số góc luôn cùng dấu với r. LSRL là đường "tốt nhất" theo nghĩa: tổng bình phương phần dư $\sum(y - \hat{y})^2$ là nhỏ nhất trong tất cả các đường thẳng có thể vẽ qua dữ liệu – đây là lý do tên gọi "bình phương tối thiểu" (least squares).

2.3.2 Interpreting slope and intercept in context

Every slope and intercept should be interpreted *in the context of the problem*, using the units of the variables involved – a bare number like "3.6" is not a complete interpretation.

Slope interpretation template: "For each additional [one unit of x], the predicted [y, in its units] [increases/decreases] by about [$|b|$], on average." **Intercept interpretation template:** "When [$x = 0$], the predicted [y] is about [a]." The intercept is only meaningful if $x = 0$ is a realistic, or at least plausible, value within (or near) the range of the data – otherwise it is simply the point where the fitted line happens to cross the y-axis, with no practical meaning.

In the study-hours example, the slope $b = 3.6$ means: for each additional hour of studying, predicted exam score increases by about 3.6 points, on average. The intercept $a = 57$ would literally mean a student who studies 0 hours is predicted to score about 57 – since $x = 0$ is not far outside the observed study-hour range (2 to 9 hours), this interpretation is at least plausible, though still an extrapolation (see below).

Interpreting slope and intercept together

A biologist records a cricket's chirp rate (x, chirps per 15 seconds) and the outdoor temperature (y, in °F) on five occasions: $(20, 60), (24, 61), (28, 73), (32, 81), (36, 90)$. Summary statistics: $\bar{x} = 28$, $s_x \approx 6.32$, $\bar{y} = 73$, $s_y \approx 12.90$, $r \approx 0.980$. Find the LSRL and interpret both the slope and the intercept in context.

$b = r \cdot \dfrac{s_y}{s_x} = 0.980 \cdot \dfrac{12.90}{6.32} \approx 2.00$. $a = \bar{y} - b\bar{x} = 73 - 2.00(28) = 73 - 56 = 17.00$. So $\hat{y} = 17.00 + 2.00x$.

Slope: for each additional chirp per 15 seconds, predicted temperature increases by about 2°F, on average.

Intercept: literally, a cricket chirping 0 times per 15 seconds would predict a temperature of about 17°F – but $x = 0$ is far below the observed chirp rates (20 to 36), so this is an extrapolation and the number 17°F should not be trusted as a real prediction (crickets typically stop chirping altogether at low temperatures well above 17°F, so the linear pattern almost certainly breaks down long before $x = 0$).

As a useful check, notice that $\hat{y}(\bar{x}) = 17 + 2(28) = 73 = \bar{y}$ exactly – the LSRL **always** passes through the point $(\bar{x}, \bar{y})$, regardless of the data set, since $a = \bar{y} - b\bar{x}$ was constructed precisely so that this holds.

2.3.3 Reading regression output from technology

In practice, the LSRL, r^2, and s are usually produced by a calculator or statistical software rather than computed by hand from summary statistics. Output is typically displayed in a table like the one below.

Reading computer regression output

Software output for the cricket-chirp data above:

Predictor	Coef
Constant	17.00
Chirps	2.00

$S = 2.944$ $\quad$ R-sq $= 96.1\%$

Write the LSRL, and find r (with the correct sign).

The row labeled "Constant" gives the intercept $a = 17.00$, and the row labeled by the explanatory variable ("Chirps") gives the slope $b = 2.00$, so $\hat{y} = 17.00 + 2.00x$ – matching the hand calculation above. $S = 2.944$ is the residual standard deviation s, and R-sq $= 96.1\%$ is $r^2 = 0.961$. Since software output reports R-sq (not r itself) and does not display a sign, recover r by taking a square root and attaching the sign of the slope: $r = +\sqrt{0.961} \approx 0.980$ (positive, since $b = 2.00 > 0$).

2.3.4 Prediction, interpolation, and the danger of extrapolation

Once a LSRL is fitted, it can be used to **predict** a value of y for a given value of x by substituting into $\hat{y} = a + bx$. Predicting for an x-value *within* the range of the observed data is called **interpolation** and is generally reliable. Predicting for an x-value *outside* the range of the observed data is called **extrapolation**, and is unreliable – there is no evidence the same linear pattern continues beyond the data that was actually collected. In the cricket example, predicting temperature at $x = 30$ chirps (within the observed range of 20 to 36) is interpolation: $\hat{y} = 17 + 2(30) = 77$, i.e. 77°F, a trustworthy prediction. Predicting at $x = 5$ chirps or $x = 60$ chirps would be extrapolation, since both fall outside the observed range.

The danger of extrapolation

Using the study-hours model $\hat{y} = 57 + 3.6x$ (fit to data ranging from about $x = 2$ to $x = 9$ hours), predict the exam score for a student who studies $x = 40$ hours, and comment on the reliability of this prediction.

$\hat{y} = 57 + 3.6(40) = 57 + 144 = 201$. This predicted score of 201 points is obviously impossible if the exam is scored out of 100 points. The value $x = 40$ is far outside the observed range of study hours (2 to 9), so this is a severe **extrapolation**: nothing in the data justifies assuming the same linear trend (roughly 3.6 points per hour) continues indefinitely – in reality, additional studying almost certainly shows *diminishing returns* well before 40 hours, and the true relationship (if it could even be measured that far out) is very unlikely to still be a straight line.

(!) Lỗi thường gặp: Ngoại suy (extrapolation) là dùng LSRL để dự đoán tại một giá trị x **nằm ngoài phạm vi dữ liệu quan sát được**. Không có gì đảm bảo xu hướng tuyến tính vẫn đúng bên ngoài phạm vi đó – dự đoán có thể vô lý (như điểm thi vượt quá thang điểm) hoặc đơn giản là sai vì mối quan hệ thực tế đổi dạng (VD không còn tuyến tính, có "diminishing returns", đạt trần, v.v.). Luôn kiểm tra: giá trị x cần dự đoán có nằm trong khoảng dữ liệu gốc hay không, trước khi tin tưởng vào $\hat{y}$.

2.4 Residuals and Residual Plots

2.4.1 The residual and the residual plot

A **residual** is the vertical distance between an observed data point and the LSRL: residual $=$ $y - \hat{y} =$ observed $-$ predicted. A positive residual means the LSRL *underpredicted* that point (the actual value was higher); a negative residual means the LSRL *overpredicted* it. For the true least-squares line, the residuals always sum to (approximately) zero. A **residual plot** graphs the residuals (vertical axis) against the explanatory variable x (horizontal axis), with a horizontal reference line at residual $= 0$.

VN **GIẢI THÍCH TIẾNG VIỆT**

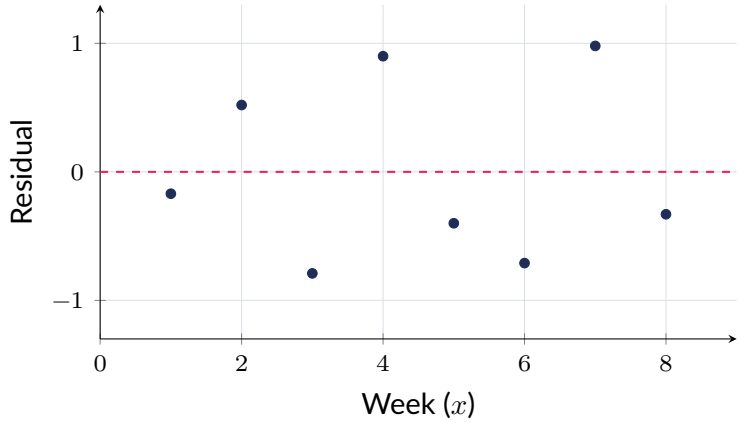
Phần dư (residual) $=$ giá trị thực tế $-$ giá trị dự đoán $= y - \hat{y}$. Phần dư dương nghĩa là LSRL **dự đoán thấp hơn** thực tế; phần dư âm nghĩa là LSRL **dự đoán cao hơn** thực tế. **Biểu đồ phần dư (residual plot)** vẽ các phần dư theo x, kèm một đường ngang tại 0 – đây là công cụ quan trọng nhất để kiểm tra xem mô hình tuyến tính có thực sự phù hợp hay không, ngay cả khi r trông có vẻ lớn.

Constructing a residual plot

A trainee's push-up count y is recorded at the end of each week $x = 1, \dots, 8$ of a training program: $(1, 5), (2, 8), (3, 9), (4, 13), (5, 14), (6, 16), (7, 20), (8, 21)$. Summary calculation gives the LSRL $\hat{y} \approx 2.857 + 2.310x$ (using $\bar{x} = 4.5$, $\bar{y} = 13.25$, $s_x \approx 2.449$, $s_y \approx 5.701$, $r \approx 0.992$, so $b = r \cdot s_y/s_x \approx 2.310$ and $a = \bar{y} - b\bar{x} \approx 2.857$). Compute the residuals and comment on whether a linear model is appropriate.

x	y (observed)	$\hat{y}$ (predicted)	residual
1	5	5.17	-0.17
2	8	7.48	0.52
3	9	9.79	-0.79
4	13	12.10	0.90
5	14	14.40	-0.40
6	16	16.71	-0.71
7	20	19.02	0.98
8	21	21.33	-0.33

The residuals bounce back and forth between roughly -0.8 and $+1.0$ with no obvious trend or curve as x increases (and they sum to 0, as expected for a true LSRL). This random, patternless scatter around zero confirms a **linear model is appropriate** for this data.

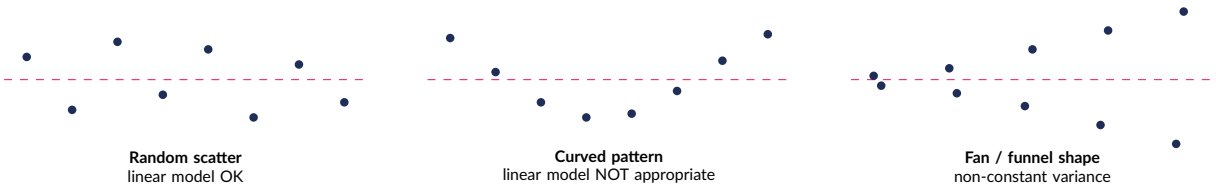


Residual plot for the push-up data: residuals scatter randomly above and below zero, with no curve or fan shape, confirming that a linear model is a reasonable fit.

2.4.2 Reading patterns in a residual plot

A residual plot's shape reveals whether the linear model is trustworthy, independent of how large r or r^2 happens to be.

Random scatter around zero, with no trend or curve ⇒ the linear model is appropriate. **Curved pattern** (e.g. a U-shape or arch) ⇒ the true relationship is *not* linear – a curve fits the data better than a line, even if r looks high. **Fan or funnel shape** (residuals spreading out, or narrowing, as x increases) ⇒ *non-constant variance*: the linear model's predictions become less (or more) precise across the range of x.



Three canonical residual-plot patterns. Left: random scatter around zero – the linear model fits well. Middle: a clear curve – the true relationship is not linear. Right: a fan (funnel) shape – the spread of residuals is not constant across x.

A high r with a clearly curved residual plot

Consider $(1,1),(2,4),(3,9),(4,16),(5,25)$ – that is, $y=x^2$. A straight-line fit gives $\hat{y}=-7+6x$ with $r\approx 0.981$ – a correlation that looks very strong. Compute the residuals and evaluate whether the linear model is really appropriate.

Predicted values: at $x=1,\dots,5$, $\hat{y}=-1,5,11,17,23$. Residuals: $1-(-1)=2$; $4-5=-1$; $9-11=-2$; $16-17=-1$; $25-23=2$. The residuals form the sequence $+2,-1,-2,-1,+2$: they start positive, dip negative in the middle, then rise back to positive – a textbook **U-shaped curve**, not random scatter. Even though $r\approx 0.981$ looks impressively strong, the residual plot reveals the true relationship is curved (in fact exactly quadratic), so a straight line is *not* the right model here. This is exactly why a high correlation alone is never sufficient evidence that a linear model is appropriate – the residual plot must also be checked.

VN **GIẢI THÍCH TIẾNG VIỆT**

Một hệ số r lớn (VD 0.98) KHÔNG đảm bảo mô hình tuyến tính là đúng – ví dụ $y=x^2$ cho $r\approx 0.981$ nhưng biểu đồ phần dư có dạng chữ U rõ rệt ($+2,-1,-2,-1,+2$), cho thấy quan hệ thực chất là parabol chứ không phải đường thẳng. Vì vậy quy trình chuẩn khi phân tích hồi quy luôn là: (1) tính LSRL và r, RỒI (2) bắt buộc vẽ và kiểm tra biểu đồ phần dư trước khi kết luận mô hình tuyến tính phù hợp.

2.5 The Coefficient of Determination r^2 and the Standard Deviation of the Residuals s

2.5.1 Interpreting r^2 in context

r^2, the **coefficient of determination**, is the proportion (or percent) of the variability in the response variable y that is explained by the least-squares regression on the explanatory variable x. It is always between 0 and 1 (0% and 100%), and it is simply the square of the correlation coefficient r.

Ví dụ mẫu · Worked example

A regression of house price (y, in \$1000s) on square footage (x) gives $\hat{y}=25+0.12x$ with $r^2=0.72$. Interpret r^2 in context.

About 72% of the variability in house price is explained by the linear relationship with square footage. The remaining 28% of the variability in price is due to other factors not captured by square footage alone (location, age, number of bathrooms, condition, and so on) – as well as ordinary scatter around the line. Note that r^2 describes the overall fit of the model across all houses in the data; it says nothing about any single house's price, and it is *not* the same number as r itself: here $r=\sqrt{0.72}\approx 0.85$, not 0.72.

VN **GIẢI THÍCH TIẾNG VIỆT**

r^2 trả lời câu hỏi: "mô hình tuyến tính giải thích được bao nhiêu phần trăm sự biến thiên của y?" – đây là một con số mô tả **độ phù hợp của cả mô hình**, không phải một dự đoán cho từng cá thể riêng lẻ. Lưu ý r^2 luôn không âm (từ 0 đến 1), khác với r có thể âm – nên khi biết r^2, ta chỉ biết **độ mạnh** của quan hệ tuyến tính, không biết được **chiều** (dương hay âm) trừ khi được cho biết thêm dấu của hệ số góc b.

2.5.2 The standard deviation of the residuals s

While r^2 describes fit as a percentage, the **standard deviation of the residuals**, s, describes the typical size of a prediction error in the *original units* of y.

$s = \sqrt{\dfrac{\sum(y-\hat{y})^2}{n-2}}$, where n is the number of data points. Roughly speaking, s is the typical ("give or take") distance between an observed y-value and the value the LSRL predicts for it – the typical prediction error.

Computing s

For the push-up training data above ($n = 8$), the residuals were $-0.17,\ 0.52,\ -0.79,\ 0.90,\ -0.40,\ -0.71,\ 0.98,\ -0.33$. Find and interpret s (and check r^2 for this same fit).

Squaring and summing: $0.17^2 + 0.52^2 + 0.79^2 + 0.90^2 + 0.40^2 + 0.71^2 + 0.98^2 + 0.33^2 \approx 3.476$. Then $s = \sqrt{\dfrac{3.476}{8-2}} = \sqrt{0.579} \approx 0.761$. Predictions from this LSRL are typically off by about 0.76 push-ups. (For reference, $r \approx 0.992$ here, so $r^2 \approx 0.985$: about 98.5% of the variability in push-up count is explained by week number – a very strong fit, consistent with the small value of s.)

VN **GIẢI THÍCH TIẾNG VIỆT**

s (độ lệch chuẩn phần dư) cho biết **sai số dự đoán điển hình**, tính bằng đúng đơn vị của y – khác với r^2 là một tỉ lệ phần trăm không đơn vị. Công thức $s = \sqrt{\sum(y-\hat{y})^2/(n-2)}$ tương tự độ lệch chuẩn thông thường nhưng chia cho $n-2$ (vì đã "dùng mất" 2 bậc tự do để ước lượng a và b). s càng nhỏ, các điểm dữ liệu càng bám sát đường LSRL, và dự đoán càng đáng tin cậy.

Where r^2 comes from. The total variation in y can be split into two pieces: variation the line explains, and variation left over as scatter around the line. Writing $SST = \sum(y-\bar{y})^2$ (total variation in y) and $SSE = \sum(y-\hat{y})^2$ (leftover variation, the sum of squared residuals), it can be shown that

$$r^2 = 1 - \frac{SSE}{SST}.$$

For the push-up data, $SSE \approx 3.476$ (computed above) and $SST = \sum(y-\bar{y})^2 \approx 227.5$ (using $\bar{y} = 13.25$). So $r^2 \approx 1 - \dfrac{3.476}{227.5} \approx 1 - 0.0153 \approx 0.9847$ – matching $r^2 \approx 0.985$ obtained directly by squaring $r \approx 0.992$. This confirms, numerically, that r^2 really is the fraction of the total variation in y that the linear model accounts for: here, only about 1.5% of the total variation in push-up counts is left unexplained (residual scatter).

2.6 Outliers and Influential Points in Regression

In the context of regression, "unusual" points come in two different, sometimes overlapping, flavors, and it is essential to tell them apart.

An **outlier** (in regression) is a point with a **large residual** – it lies far from the pattern followed by the rest of the data, vertically. An **influential point** is a point whose **removal would substantially change the LSRL** (its slope and/or intercept) – influential points typically have an **extreme x-value** (far from $\bar{x}$), giving them high *leverage* to pull the line toward themselves. A point can be an outlier, influential, both, or neither – and, surprisingly, an influential point does not need to have a large residual once it is included in the fit.

A large-residual point that is not influential

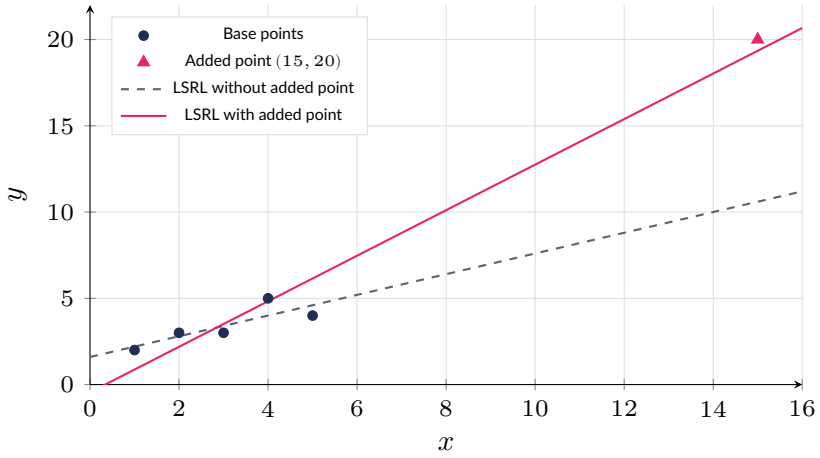
Base data: $(1,2),(2,3),(3,3),(4,5),(5,4)$, giving $\bar{x}=3$, LSRL $\hat{y}=1.6+0.6x$, $r\approx 0.832$. Suppose a sixth point, $(3,12)$, is added – note its x-value, 3, equals $\bar{x}$ exactly. Find the new LSRL and comment.

Recomputing with all six points: the new mean is still $\bar{x}=3$ (since $\frac{5(3)+3}{6}=3$), so the added point's deviation from the mean, $x-\bar{x}=0$, contributes *nothing* to the slope calculation ($b=\sum(x-\bar{x})(y-\bar{y})/\sum(x-\bar{x})^2$ – a zero x-deviation multiplies every term involving this point by 0). The new LSRL is $\hat{y}\approx 3.033+0.6x$: the slope is **unchanged** at 0.6, only the intercept shifted (from 1.6 to about 3.03) and r dropped sharply (from 0.832 to about 0.232, since the point clashes badly with the pattern). The predicted value at $x=3$ is now about 4.83, so this point's residual is $12-4.83\approx 7.17$ – a huge **outlier** by residual – yet it barely influenced the slope at all, because a point sitting exactly at $\bar{x}$ has *zero leverage*.

An influential point with high leverage

Using the same base data $(1,2),(2,3),(3,3),(4,5),(5,4)$ ($\hat{y}=1.6+0.6x$), suppose instead a point far out at $x=15$ is added: $(15,20)$. Find the new LSRL and comment.

Recomputing with this point included: $\bar{x}=5$, $\bar{y}\approx 6.17$, and the new LSRL is $\hat{y}\approx -0.45+1.32x$ (with $r\approx 0.984$). The slope more than **doubled**, from 0.6 to about 1.32 – a dramatic change caused by a single point. Checking this point's own residual under the *new* line: predicted value at $x=15$ is $-0.45+1.32(15)\approx 19.40$, so the residual is only $20-19.40\approx 0.60$ – quite small! This point does *not* stand out as an outlier once it is included (the line has been pulled toward it), yet it is clearly **influential**: removing it swings the slope back down to 0.6. The key giveaway is its extreme x-value (15, far from the other points' range of 1 to 5), which gives it strong leverage over the fitted line.



Adding a single point with an extreme x-value tilts the LSRL substantially (dashed vs. solid pink line): a clear case of an influential point.

(!) Lỗi thường gặp: Đừng đánh đồng "điểm ngoại lệ (outlier)" với "điểm ảnh hưởng (influential point)". **Outlier** chỉ nói về **phần dư lớn** (điểm nằm xa mô hình theo phương thẳng đứng) – một outlier ở gần $\bar{x}$ có thể gần như **không ảnh hưởng** đến hệ số góc. **Influential point** nói về việc **loại bỏ điểm đó làm thay đổi đáng kể LSRL** – thường do điểm có giá trị x cực đoan (đòn bẩy/leverage cao), và một điểm có thể ảnh hưởng mạnh dù phần dư của nó (sau khi đã tính cả nó vào mô hình) lại khá nhỏ. Cách kiểm tra chắc chắn nhất: tính lại LSRL **có và không có** điểm

nghi ngờ, rồi so sánh hệ số góc/hệ số chặn.

Feature	Outlier	Influential point
Defined by	a large residual (far from the pattern vertically)	a substantial change in the LSRL's slope/intercept when the point is removed
Usual cause	a y-value inconsistent with the rest of the data	an extreme x-value (far from $\bar{x}$), giving high leverage
Point at $x = \bar{x}$	can still be a large-residual outlier	has essentially *zero* leverage – cannot be influential on the slope
Residual once included	typically large	can be small (the line may already be pulled toward the point)
How to confirm	inspect the residual plot for a point far from 0	refit the LSRL with and without the point; compare a and b

Putting the two ideas together, every unusual point falls into one of four categories: a **typical point** (neither an outlier nor influential – it fits the pattern and sits near $\bar{x}$); an **outlier only** (large residual, but near $\bar{x}$, like the $(3, 12)$ point above – low leverage means the slope barely moves); an **influential point only** (extreme x-value that pulls the line toward itself, ending up with a small residual once included, like the $(15, 20)$ point above); and a point that is **both** an outlier and influential (extreme x-value *and* a large residual even after the line is refit) – this last case is the most damaging, since the point both distorts the LSRL substantially and remains poorly fit by the resulting line.

VN **GIẢI THÍCH TIẾNG VIỆT**

Bốn khả năng khi có một điểm bất thường trong hồi quy: (1) **điểm bình thường** – không phải outlier, không ảnh hưởng; (2) **chỉ là outlier** – phần dư lớn nhưng x gần $\bar{x}$ nên hầu như không đổi hệ số góc; (3) **chỉ là influential** – x cực đoan nên "kéo" đường hồi quy về phía nó, và sau khi tính vào mô hình thì phần dư của chính nó lại nhỏ; (4) **vừa là outlier vừa influential** – x cực đoan VÀ phần dư vẫn lớn ngay cả sau khi đã tính vào mô hình – đây là trường hợp đáng lo ngại nhất.

2.7 Transforming Data to Achieve Linearity

2.7.1 Exponential growth and the logarithmic transformation

Not every relationship between two quantitative variables is linear. When a scatterplot of y versus x curves upward at an increasing rate – consistent with **exponential growth**, $y = a \cdot b^x$ – a logarithmic transformation can straighten the pattern so that least-squares regression can be applied.

If $y = a \cdot b^x$ (an exponential model), then taking the natural log of both sides gives $\ln y = \ln a + x \ln b$ – a **linear** equation in x and $\ln y$. So: (1) transform the data by computing $\ln y$ for each point; (2) fit the LSRL of $\ln y$ on x, giving $\widehat{\ln y} = a_1 + b_1 x$; (3) if the transformed scatterplot is linear (and its residual plot shows random scatter), the exponential model fits, and translating back: $a = e^{a_1}$, $b = e^{b_1}$, giving $\hat{y} = a \cdot b^x = e^{a_1} \cdot \left(e^{b_1}\right)^x$.

VN **GIẢI THÍCH TIẾNG VIỆT**

Khi biểu đồ phân tán y theo x cong lên nhanh dần (tăng trưởng mũ, exponential), ta không thể áp dụng LSRL trực tiếp. Giải pháp: lấy $\ln y$ (log tự nhiên), vẽ lại biểu đồ phân tán của $\ln y$ theo x – nếu biểu đồ mới thẳng hàng, mô hình mũ $y = a \cdot b^x$ phù hợp. Sau khi hồi quy $\ln y$ theo x được $\widehat{\ln y} = a_1 + b_1 x$, "dịch ngược" lại bằng cách lấy mũ hai vế: $a = e^{a_1}$ (giá trị ban đầu, ứng với $x = 0$) và $b = e^{b_1}$ (hệ số tăng trưởng mỗi đơn vị x).

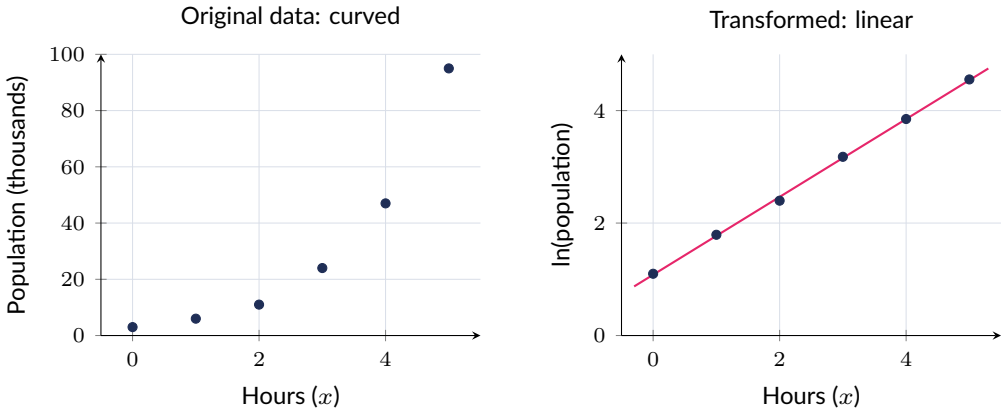
Linearizing exponential growth

A biologist counts a bacterial colony's population y (in thousands) at the end of each hour $x = 0, 1, 2, 3, 4, 5$: $y = 3,\ 6,\ 11,\ 24,\ 47,\ 95$. The scatterplot of y vs. x curves sharply upward. Apply a log transformation, fit a line, and translate back to an exponential model.

Step 1 – transform: $\ln y = 1.0986,\ 1.7918,\ 2.3979,\ 3.1781,\ 3.8501,\ 4.5539$ for $x = 0, \dots, 5$ respectively.
Step 2 – fit the LSRL of $\ln y$ on x: using $\bar{x} = 2.5$, $\overline{\ln y} \approx 2.812$, $s_x \approx 1.871$, $s_{\ln y} \approx 1.296$, and $r \approx 0.99997$ (essentially perfect – confirming the transformed data is very linear), the slope is $b_1 \approx 0.692$ and intercept $a_1 \approx 1.081$, so $\widehat{\ln y} = 1.081 + 0.692x$.
Step 3 – translate back: $a = e^{1.081} \approx 2.95$ and $b = e^{0.692} \approx 2.00$, giving the exponential model

$$\hat{y} \approx 2.95 \cdot (2.00)^x.$$

Since $b \approx 2.00$, the model says the colony's population roughly **doubles every hour** – consistent with the data (population values are close to doubling at each step). Using the model to predict at $x = 6$: $\hat{y} \approx 2.95(2.00)^6 \approx 2.95(64) \approx 188$ thousand.



Left: the original population data curves sharply upward. Right: plotting $\ln y$ against x straightens the pattern almost perfectly, confirming an exponential model.

After fitting a transformed regression, it is still important to check the residual plot of the transformed fit (residuals of $\ln y$ versus x) for random scatter – the same diagnostic used throughout this unit, now applied on the transformed scale. For the bacterial-growth data, the residuals of $\widehat{\ln y} = 1.081 + 0.692x$ are $0.018,\ 0.019,\ -0.068,\ 0.020,\ -0.000,\ 0.011$ for $x = 0, \dots, 5$: all small, with no systematic curve, confirming that the exponential model – not merely a coincidentally high r – genuinely fits this data.

2.7.2 Power models and the log-log transformation

A second common curved pattern is the **power model**, $y = a \cdot x^p$ (for $x > 0$), which arises in contexts like scaling laws (e.g. an animal's metabolic rate as a power of its body mass). Power models are linearized differently than exponential models: instead of transforming only y, *both* variables are log-transformed.

> If $y = a \cdot x^p$, then $\ln y = \ln a + p \ln x$ – linear in $\ln x$ and $\ln y$. So: plot $\ln y$ versus $\ln x$ (a **log-log** transformation); if that scatterplot is linear, fit $\widehat{\ln y} = a_1 + b_1 \ln x$. Then the exponent is simply $p = b_1$ (the slope), and $a = e^{a_1}$, giving $\hat{y} = a \cdot x^p$.

Recovering a power model from log-log regression

A log-log regression (base-10 logs) on a data set gives $\log \hat{y} = 0.30 + 0.50 \log x$. Identify the power model $\hat{y} = a \cdot x^p$.

Here the slope is $p = 0.50$ and the intercept converts as $a = 10^{0.30} \approx 1.995 \approx 2.00$. So the fitted power model is $\hat{y} \approx 2.00 \cdot x^{0.5} = 2.00\sqrt{x}$ – a square-root relationship between x and y.

Reasoning: exponential or power model?

Two researchers each have a curved scatterplot. Researcher A's data starts near $y = 0$ when x is near 0 and grows at a fairly steady *percentage* rate as x increases (each additional unit of x multiplies y by roughly the same factor). Researcher B's data is only defined for $x > 0$, passes near the origin, and grows more slowly in percentage terms as x gets large (each doubling of x multiplies y by roughly the same factor, rather than each added unit of x). Which transformation – $\ln y$ only, or log-log – is more appropriate for each?

Researcher A: growth by a constant *multiplicative factor per added unit of* x is the signature of **exponential** growth ($y = a \cdot b^x$), so only $\ln y$ should be transformed (plot $\ln y$ vs. x).
Researcher B: growth that depends on *ratios* of x (each doubling of x multiplies y by a fixed factor) is the signature of a **power** model ($y = a \cdot x^p$), which calls for the log-log transformation (plot $\ln y$ vs. $\ln x$).

VN **GIẢI THÍCH TIẾNG VIỆT**

Mô hình lũy thừa (power model) $y = a \cdot x^p$ khác mô hình mũ ở chỗ số mũ nằm ở x, không phải ở hằng số. Để tuyến tính hoá, phải lấy log ở **CẢ HAI** biến ($\ln x$ và $\ln y$), không chỉ y như mô hình mũ. Sau khi hồi quy $\ln y$ theo $\ln x$ được hệ số góc b_1, đó chính là số mũ p của mô hình gốc; hệ số chặn a_1 chuyển ngược lại bằng $a = e^{a_1}$ (hoặc 10^{a_1} nếu dùng log cơ số 10).

Feature	Exponential model $y = a \cdot b^x$	Power model $y = a \cdot x^p$
Raw scatterplot	curves upward (or decays) faster and faster; multiplicative growth per unit x	curves following a power law, typically defined for $x > 0$
Transform applied	$\ln y$ only (x left alone)	both $\ln y$ *and* $\ln x$ (log-log)
Linearized equation	$\ln y = \ln a + x \ln b$	$\ln y = \ln a + p \ln x$
Recovering the model	$a = e^{a_1},\ b = e^{b_1}$	$a = e^{a_1},\ p = b_1$ (the slope itself)

Key formulas for this unit. Marginal distribution: row/column total ÷ grand total. **Conditional distribution:** cell count ÷ that row's/column's total.

Correlation: $r = \dfrac{1}{n-1} \sum \left(\dfrac{x - \bar{x}}{s_x} \right) \left(\dfrac{y - \bar{y}}{s_y} \right), -1 \le r \le 1.$

LSRL: $\hat{y} = a + bx,\ b = r \cdot \dfrac{s_y}{s_x},\ a = \bar{y} - b\bar{x}.$

Residual: $y - \hat{y}$. **Coefficient of determination:** $r^2 = 1 - \dfrac{SSE}{SST}$. **Residual standard deviation:** $s = \sqrt{\dfrac{\sum (y - \hat{y})^2}{n-2}}$.

Exponential transform: $y = a \cdot b^x \iff \ln y = \ln a + x \ln b$. **Power transform:** $y = a \cdot x^p \iff \ln y = \ln a + p \ln x$.

2.8 Practice Problems

Bài tập luyện tập

For a bivariate data set, $\bar{x} = 12$, $s_x = 3$, $\bar{y} = 50$, $s_y = 12$, and $r = -0.75$. What is the LSRL's slope, and what does the sign mean?

(A) $b = -3$; as x increases, y tends to decrease

(B) $b = -3$; as x increases, y tends to increase

(C) $b = 3$; as x increases, y tends to decrease

(D) $b = 0.75$; no relationship

Lời giải chi tiết

$b = r \cdot \dfrac{s_y}{s_x} = -0.75 \cdot \dfrac{12}{3} = -0.75(4) = -3$. Since the slope is negative, y tends to **decrease** as x increases. **(A)**.

Bài tập luyện tập

A regression of house price (y, in \$1000s) on square footage (x) gives $\hat{y} = 25 + 0.12x$ with $r^2 = 0.72$. Which interpretation is correct?

(A) 72% of houses are priced above average

(B) For each extra square foot, price increases by exactly \$0.12

(C) About 72% of the variation in price is explained by the linear model on square footage

(D) The correlation is $r = 0.72$

Lời giải chi tiết

r^2 is the proportion of variability in the response (y) explained by the least-squares regression on x; it is a description of the model's fit, not a statement about individual houses. **(C)**. (Note: $r = \sqrt{0.72} \approx 0.85$, not 0.72 – common trap in choice (D).)

Bài tập luyện tập

In the study-hours example above ($\hat{y} = 57 + 3.6x$), a student who studied $x = 8$ hours actually scored $y = 86$. Find and interpret the residual.

Lời giải chi tiết

Predicted: $\hat{y} = 57 + 3.6(8) = 57 + 28.8 = 85.8$. Residual $= y - \hat{y} = 86 - 85.8 = 0.2$. The residual is small and positive: the LSRL predicted almost exactly right, slightly underestimating this student's score by 0.2 points.

Bài tập luyện tập

A survey of 400 smartphone owners records operating system (iOS or Android) and satisfaction level (Satisfied, Neutral, Dissatisfied):

	Satisfied	Neutral	Dissatisfied	Total
iOS	140	40	20	200
Android	90	60	50	200
Total	230	100	70	400

What percent of ALL 400 respondents are iOS users who reported being Satisfied?

(A) 35%
(B) 70%
(C) 57.5%
(D) 50%

Lời giải chi tiết

This asks for a **joint** percentage (both conditions at once, out of the grand total): $\frac{140}{400} = 35\%$. (Choice (B), 70%, is instead the *conditional* percent of iOS users who are satisfied, $140/200$ – a common mix-up between joint and conditional distributions.) **(A)**.

Bài tập luyện tập

Using the table above, find the marginal distribution of satisfaction level, and the conditional distribution of satisfaction given Android. Is there an association between operating system and satisfaction? Explain using the conditional distributions.

Lời giải chi tiết

Marginal distribution of satisfaction (divide column totals by the grand total 400): Satisfied $= 230/400 = 57.5\%$, Neutral $= 100/400 = 25\%$, Dissatisfied $= 70/400 = 17.5\%$.
Conditional distribution given Android (divide the Android row by its total, 200): Satisfied $= 90/200 = 45\%$, Neutral $= 60/200 = 30\%$, Dissatisfied $= 50/200 = 25\%$.
Conditional distribution given iOS (for comparison, divide the iOS row by 200): Satisfied $= 140/200 = 70\%$, Neutral $= 40/200 = 20\%$, Dissatisfied $= 20/200 = 10\%$.
Because the conditional distributions differ substantially between iOS ($70/20/10$) and Android ($45/30/25$), there **is an association**: iOS owners report being satisfied at a much higher rate (70%) than Android owners (45%).

Bài tập luyện tập

To assess whether there is an association between two categorical variables displayed in a two-way table, the correct approach is to:

(A) Compare the two marginal distributions to each other
(B) Compare the conditional distributions of one variable across the categories of the other
(C) Compare the grand total to each individual cell count
(D) Compute the correlation coefficient r between the two variables

Lời giải chi tiết

Association is assessed by comparing how the distribution of one variable *changes* across categories of the other – i.e., by comparing conditional distributions (not marginal distributions,

which by definition ignore the other variable entirely; and r applies to quantitative variables, not categorical ones). **(B)**.

Bài tập luyện tập

In the segmented bar chart of transportation mode by class year shown earlier in this unit, which statement is best supported by the data?

(A) Underclassmen and upperclassmen use each mode of transportation at roughly the same rate

(B) A majority of underclassmen drive to campus

(C) A majority of upperclassmen drive to campus (70%), while walking is the most common mode among underclassmen (45%)

(D) Bus is the most common mode of transportation for both groups

Lời giải chi tiết

From the conditional distributions computed earlier: Underclassmen walk most often (45%, the largest of their three categories, though not a majority), while 70% of Upperclassmen – a true majority – drive. This is a clear association, ruling out (A). Choice (D) is false (Bus is the least common mode for both groups). **(C)**.

Bài tập luyện tập

A scatterplot of daily hours spent watching TV (x) and GPA (y) for a group of students shows points loosely scattered, trending downward from upper left to lower right, with considerable scatter around the overall pattern. Which value of r is most plausible?

(A) $r = 0.85$

(B) $r = -0.85$

(C) $r = -0.35$

(D) $r = 0.35$

Lời giải chi tiết

The trend is downward, so r must be **negative**, eliminating (A) and (D). "Considerable scatter" around the pattern (not a tight, clean line) rules out a value close to -1, so a modest magnitude like $r = -0.35$ is far more plausible than $r = -0.85$. **(C)**.

Bài tập luyện tập

Suppose the correlation between adults' height (measured in centimeters) and weight (measured in kilograms) is $r = 0.78$. (a) If height were instead measured in inches and weight in pounds, what would the new correlation be? (b) If the variables' roles were swapped, with $x =$ weight and $y =$ height, what would r be? Justify both answers.

Lời giải chi tiết

(a) $r = 0.78$ still. Correlation is **unaffected by a change of units** (converting cm to inches and kg to pounds is a linear rescaling of each variable, and r is unitless).
(b) $r = 0.78$ still. Correlation is **symmetric in x and y** – it measures the strength of the linear association between the two variables regardless of which one is labeled explanatory (x) and which is labeled response (y).

Bài tập luyện tập

Four points follow a perfect line: $(1, 5), (2, 10), (3, 15), (4, 20)$, giving $r = 1$. A fifth point, $(5, 2)$, is added. Find the new value of r (using $b = r \cdot s_y/s_x$ and the fact that recomputing gives slope $b = 0.4$ and intercept $a = 9.2$ for the five-point data set; you do not need to derive b and a from scratch) and comment on what this illustrates about r.

Lời giải chi tiết

With the new point included, recomputing the correlation for all five points gives $r \approx 0.087$ – essentially 0. A single point added far from an otherwise perfect linear pattern collapsed the correlation from $r = 1$ (perfect positive) to $r \approx 0.087$ (essentially no linear association). This is a dramatic illustration that r is **not resistant** to outliers, especially in small data sets.

Bài tập luyện tập

Summary statistics for advertising spend (x, in \$1000s) and sales (y, in \$1000s): $\bar{x} = 20$, $s_x = 4$, $\bar{y} = 150$, $s_y = 30$, $r = 0.6$. Find the LSRL.

Lời giải chi tiết

$b = r \cdot \dfrac{s_y}{s_x} = 0.6 \cdot \dfrac{30}{4} = 0.6(7.5) = 4.5$. $a = \bar{y} - b\bar{x} = 150 - 4.5(20) = 150 - 90 = 60$. So $\hat{y} = 60 + 4.5x$.

Bài tập luyện tập

Using $\hat{y} = 60+4.5x$ from the previous problem (advertising spend x vs. sales y, both in \$1000s), which is the correct interpretation of the slope?

(A) For each additional dollar spent on advertising, sales increase by \$4.5

(B) For each additional \$1000 spent on advertising, predicted sales increase by about \$4500, on average

(C) When advertising spend is \$0, predicted sales are \$4500

(D) Sales cause advertising spend to increase by \$4500

Lời giải chi tiết

Both variables are measured in \$1000s, so the slope 4.5 means: for each additional \$1000 (one unit of x) spent on advertising, predicted sales increase by about 4.5 thousand dollars (\$4500), on average. Choice (A) misreads the units; (C) confuses slope with intercept; (D) wrongly asserts causation and reverses the variables. **(B)**.

Bài tập luyện tập

For the same model $\hat{y} = 60 + 4.5x$, the observed advertising spends in the data ranged from $x = 10$ to $x = 50$ (i.e., \$10,000 to \$50,000). Interpret the intercept $a = 60$ literally, and comment on whether that literal interpretation is trustworthy.

Lời giải chi tiết

Literally, the intercept says: when advertising spend is \$0, predicted sales are \$60,000. However, $x = 0$ lies **outside** the observed range of the data ($\$10{,}000$ to $\$50{,}000$), so this interpretation is an **extrapolation** – there is no evidence the same linear relationship continues down to $x = 0$ (a company spending nothing on advertising may behave very differently than the linear trend among $\$10{,}000$–$\$50{,}000$ spenders suggests). The literal number should not be trusted as a reliable prediction.

Bài tập luyện tập

Using $\hat{y} = 60 + 4.5x$ (advertising spend x in \$1000s, observed range $\$10{,}000$ to $\$50{,}000$), predict sales when advertising spend is $\$35{,}000$.

Lời giải chi tiết

$x = 35$: $\hat{y} = 60 + 4.5(35) = 60 + 157.5 = 217.5$, i.e., predicted sales of about $\$217{,}500$. Since $x = 35$ is within the observed range (10 to 50), this is interpolation, not extrapolation, so the prediction is reasonably trustworthy.

Bài tập luyện tập

For the advertising model above, with observed x ranging from 10 to 50, which of the following predictions would involve extrapolation?

(A) $x = 15$
(B) $x = 25$
(C) $x = 45$
(D) $x = 80$

Lời giải chi tiết

Extrapolation means predicting outside the range of observed x-values (10 to 50). Only $x = 80$ falls outside this range; the other three values are all within it. **(D)**.

Bài tập luyện tập

Using $\hat{y} = 60 + 4.5x$, an advertising spend of $x = 35$ (\$35,000) actually produced sales of $y = 230$ (\$230,000). Find the residual.

(A) 12.5
(B) -12.5
(C) 217.5
(D) 230

Lời giải chi tiết

$\hat{y} = 60 + 4.5(35) = 217.5$. Residual $= y - \hat{y} = 230 - 217.5 = 12.5$. **(A)**.

Bài tập luyện tập

A residual plot shows points randomly scattered above and below the zero line, with no obvious pattern, across the entire range of x. What does this indicate?

(A) The linear model is appropriate

(B) The linear model is not appropriate; the relationship is curved

(C) The residuals show non-constant variance

(D) There is a strong correlation between the variables

Lời giải chi tiết

Random, patternless scatter around zero in a residual plot is precisely the condition that indicates a linear model is a good fit for the data. **(A)**.

Bài tập luyện tập

A residual plot for six data points ($x = 1$ through 6) has these residuals, in order: $-4, -1, 1, 2, 1, -4$. Describe the pattern and state its implication for the appropriateness of the linear model.

Lời giải chi tiết

The residuals rise from negative (-4) up to positive ($+2$ at the middle) and then fall back to negative (-4) – an arch shape (concave down), a clear **curved pattern** rather than random scatter. This indicates the true relationship between x and y is **not linear**; a curved model would fit the data better than a straight line, regardless of how large r might be for the straight-line fit.

Bài tập luyện tập

A residual plot shows residuals clustered very close to zero for small values of x, but increasingly spread out (both above and below zero) as x increases. What does this pattern indicate?

(A) The linear model is appropriate and no further action is needed

(B) The relationship between x and y is curved, not linear

(C) Non-constant variance (a fan/funnel shape): predictions become less precise as x increases

(D) A strong negative correlation between x and y

Lời giải chi tiết

Residuals that fan out (or narrow) as x changes indicate the spread of the residuals is not the same across the range of x – a violation of constant variance, commonly called a fan or funnel shape, meaning predictions are less reliable in the region where residuals spread out more. **(C)**.

Bài tập luyện tập

The correlation between weekly hours of exercise (x) and reduction in resting heart rate (y, in bpm) is $r = 0.88$. Find and interpret r^2.

Lời giải chi tiết

$r^2 = 0.88^2 = 0.7744$, or about 77.4%. About 77.4% of the variability in resting heart rate reduction is explained by the linear relationship with weekly hours of exercise; the remaining 22.6% is due to other factors and natural scatter.

Bài tập luyện tập

For a data set with $n = 10$ points, the sum of squared residuals is 72. Find s, the standard deviation of the residuals.

(A) 3
(C) 7.2
(B) 9
(D) 8

Lời giải chi tiết

$s = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{72}{10-2}} = \sqrt{\frac{72}{8}} = \sqrt{9} = 3$. **(A)**.

Bài tập luyện tập

Which of the following best describes an **influential point** in a regression?

(A) Any point with a large residual
(C) A point that lies exactly on the LSRL
(B) A point whose removal would substantially change the slope or intercept of the LSRL, typically due to an extreme x-value
(D) The single point in the data set with the largest y-value

Lời giải chi tiết

An influential point is defined by its *effect on the regression line if removed*, not by the size of its residual – influential points are typically extreme in x (high leverage), and can even have a small residual once included in the fit. **(B)**.

Bài tập luyện tập

Base data: $(2,5),(4,8),(6,10),(8,15),(10,17)$ gives LSRL $\hat{y} \approx 1.7 + 1.55x$ ($r \approx 0.990$). A sixth point, $(30,10)$, is added; recomputing gives the new LSRL $\hat{y} \approx 10.03 + 0.081x$ ($r \approx 0.185$). (a) By how much does the slope change? (b) Find the residual of the added point $(30,10)$ using the *new* LSRL, and use it, together with part (a), to argue this point is influential.

Lời giải chi tiết

(a) The slope drops from about 1.55 to about 0.081 – a change of about 1.47, essentially flattening the line (more than a 90% reduction). This alone shows the point is highly **influential**.
(b) Predicted value at $x = 30$ using the new line: $\hat{y} \approx 10.03 + 0.081(30) \approx 10.03 + 2.43 \approx 12.45$. Residual $= 10 - 12.45 \approx -2.45$ – moderate, not enormous, given the response values in this data set range up to about 17. So the point's own fit is not wildly out of line, yet its presence (due to its extreme $x = 30$, far beyond the base range of 2 to 10) drastically reshaped the entire regression – confirming it is influential regardless of how "normal" its residual looks.

Bài tập luyện tập

A scatterplot of y versus x curves sharply upward, increasing faster and faster as x increases. A scatterplot of $\ln y$ versus x, however, is very linear. Which model best describes the original relationship between x and y?

(A) Linear model, $y = a + bx$
(B) Exponential model, $y = a \cdot b^x$
(C) Power model, $y = a \cdot x^p$
(D) No relationship exists

Lời giải chi tiết

A relationship that is linearized by taking $\ln y$ (leaving x untransformed) is the signature of an **exponential** model, since $y = a \cdot b^x \iff \ln y = \ln a + x \ln b$, linear in x and $\ln y$. A power model would instead require transforming *both* variables (log-log). **(B)**.

Bài tập luyện tập

An investment's value y (in dollars) is recorded at the end of each year $x = 0, 1, 2, 3, 4$ (with $x = 0$ the initial investment): $y = 1000, 1210, 1460, 1770, 2140$. The scatterplot curves upward. Use a log transformation to fit an exponential model, and predict the investment's value after 6 years.

Lời giải chi tiết

Taking natural logs: $\ln y \approx 6.9078, 7.0984, 7.2862, 7.4787, 7.6686$ for $x = 0, \ldots, 4$. Fitting the LSRL of $\ln y$ on x gives (using $\bar{x} = 2$, $\overline{\ln y} \approx 7.288$, and $r \approx 0.99999$, confirming an excellent linear fit) slope $b_1 \approx 0.190$ and intercept $a_1 \approx 6.908$, so $\widehat{\ln y} = 6.908 + 0.190x$.
Translating back: $a = e^{6.908} \approx 1000$ and $b = e^{0.190} \approx 1.21$, giving $\hat{y} \approx 1000 \cdot (1.21)^x$ – consistent with an investment growing about 21% per year. Predicting at $x = 6$: $\hat{y} \approx 1000(1.21)^6 \approx 1000(3.1298) \approx \$3130$ (extrapolating two years beyond the observed data, so this prediction should be treated with some caution, though modest here since it is only slightly beyond the observed range).

Bài tập luyện tập

To linearize a suspected **power model** $y = a \cdot x^p$ (with $x > 0$, $y > 0$), which transformation should be applied before fitting a LSRL?

(A) Plot $\ln y$ versus x
(B) Plot y versus $\ln x$
(C) Plot $\ln y$ versus $\ln x$
(D) Plot y^2 versus x^2

Lời giải chi tiết

Since $y = a \cdot x^p \iff \ln y = \ln a + p \ln x$, both variables must be log-transformed (a log-log plot) to produce a linear relationship, with the resulting slope equal to the power p. **(C)**.

Bài tập luyện tập

Six data points: $(1, 4), (2, 7), (3, 9), (4, 14), (5, 17), (6, 19)$. (a) Find the LSRL. (b) Find r and r^2. (c) Find the residual at $x = 3$. (d) Even though r is very high here, explain why a residual plot should still be examined before concluding the linear model is appropriate.

Lời giải chi tiết

(a) Using $\bar{x} = 3.5$, $\bar{y} \approx 11.67$, $s_x \approx 1.871$, $s_y \approx 5.680$, and $r \approx 0.993$: $b = r \cdot s_y / s_x \approx 3.14$, $a = \bar{y} - b\bar{x} \approx 11.67 - 3.14(3.5) \approx 0.67$. So $\hat{y} \approx 0.67 + 3.14x$.
(b) $r \approx 0.993$, so $r^2 \approx 0.986$: about 98.6% of the variability in y is explained by the linear model

on x.
(c) Predicted at $x = 3$: $\hat{y} \approx 0.67 + 3.14(3) = 0.67 + 9.42 \approx 10.10$. Residual $= 9 - 10.10 \approx -1.10$.
(d) A high r (or r^2) only measures the strength of a *linear* association – it does not rule out systematic curvature. As shown earlier in this unit, data that is exactly quadratic ($y = x^2$) can still produce $r \approx 0.98$, yet its residual plot reveals an unmistakable curved pattern. Only the residual plot can confirm that the scatter around the line is truly random (rather than systematically curved or fanning out), which is why checking it is a required step even when r looks excellent.

Bài tập luyện tập

Find r for $(1,1), (2,3), (3,2), (4,5), (5,4)$ using standardized values (z-scores). (You may use $\bar{x} = 3$, $\bar{y} = 3$, $s_x = s_y \approx 1.581$.)

Lời giải chi tiết

Standardizing: $z_x = -1.265, -0.632, 0, 0.632, 1.265$ and, matching each point in order, $z_y = -1.265, 0, -0.632, 1.265, 0.632$. Products $z_x z_y$: $(-1.265)(-1.265) = 1.600$; $(-0.632)(0) = 0.000$; $(0)(-0.632) = 0.000$; $(0.632)(1.265) = 0.800$; $(1.265)(0.632) = 0.800$. Summing: $1.600 + 0.000 + 0.000 + 0.800 + 0.800 = 3.200$. So $r = \dfrac{3.200}{5-1} = \dfrac{3.200}{4} = 0.800$.

Bài tập luyện tập

Across a large collection of countries, per-capita chocolate consumption and the number of Nobel laureates per capita are found to be strongly positively correlated. Does this mean eating chocolate causes Nobel-worthy achievement? Suggest a plausible lurking variable.

Lời giải chi tiết

No – this association almost certainly does not reflect a direct causal effect of chocolate. A plausible lurking variable is a country's **wealth (GDP per capita)**: wealthier countries tend to have both higher average chocolate consumption (a discretionary, relatively expensive food in many regions) *and* greater investment in education, universities, and scientific research (which produces more Nobel laureates) – national wealth drives both variables independently, producing a strong correlation between them with no causal link running directly from chocolate to Nobel prizes.

Bài tập luyện tập

For a regression with $n = 12$ data points, $SST = \sum(y - \bar{y})^2 = 500$ and $SSE = \sum(y - \hat{y})^2 = 120$. Find (a) r^2 and (b) s, the standard deviation of the residuals.

Lời giải chi tiết

(a) $r^2 = 1 - \dfrac{SSE}{SST} = 1 - \dfrac{120}{500} = 1 - 0.24 = 0.76$, so about 76% of the variability in y is explained by the linear model.
(b) $s = \sqrt{\dfrac{SSE}{n-2}} = \sqrt{\dfrac{120}{12-2}} = \sqrt{\dfrac{120}{10}} = \sqrt{12} \approx 3.46$.

Collecting Data

Mục tiêu Unit: Phân biệt tổng thể (population), mẫu (sample) và khung mẫu (sampling frame); thống kê mẫu (statistic) và tham số tổng thể (parameter); các phương pháp chọn mẫu ngẫu nhiên (SRS, phân tầng, cụm, hệ thống) và cách thực hiện cụ thể; các nguồn sai lệch (bias) trong khảo sát, kể cả hiệu ứng cách đặt câu hỏi; ngôn ngữ và nguyên tắc thiết kế thí nghiệm (biến giải thích, biến đáp ứng, nghiệm thức, đối tượng thí nghiệm, nhân tố, mức độ); nguyên tắc so sánh, ngẫu nhiên hoá, lặp lại; thiết kế khối ngẫu nhiên và cặp ghép; nhiễu biến (confounding), thiết kế mù đơn/mù đôi và hiệu ứng giả dược; phạm vi suy luận (scope of inference) theo khung 2×2; và các nguyên tắc đạo đức khi thu thập dữ liệu.

3.1 Planning a Study: Population, Sample, and Sampling Frame

Every statistical investigation begins with a precise description of who or what is being studied and how information about them will actually be gathered. Getting this vocabulary exactly right matters, because the validity of every later conclusion depends on knowing exactly which group a set of data can legitimately describe.

Population, sample, and sampling frame

The **population** is the entire group of individuals or objects about which information is wanted. A **sample** is the subset of the population that is actually observed or measured. The **sampling frame** is the list (or other mechanism) from which the sample is physically drawn. Ideally the sampling frame matches the population exactly, but in practice it is often a step removed from it — a phone directory is a sampling frame for "all households," but it misses unlisted numbers; a list of registered voters is a sampling frame for "eligible voters," but it misses eligible people who never registered. A **census** attempts to collect data from *every* member of the population; a **sample survey** collects data from a sample and uses it to draw conclusions about the larger population.

A **statistic** is a number that describes a *sample* (e.g., a sample mean $\bar{x}$ or a sample proportion $\hat{p}$) — it can be calculated directly from data. A **parameter** is a number that describes the *population* (e.g., a population mean μ or a population proportion p) — it is a fixed but almost always unknown number. Estimating an unknown parameter using a statistic computed from a sample, and describing how much that statistic could vary from sample to sample, is the central task of statistical inference, formally developed in the sampling-distributions and inference units.

VN GIẢI THÍCH TIẾNG VIỆT

Tổng thể (population) là toàn bộ nhóm đối tượng mà ta muốn tìm hiểu; **mẫu** (sample) là tập con thực sự được quan sát; **khung mẫu** (sampling frame) là danh sách cụ thể dùng để rút mẫu ra — trên lý thuyết khung mẫu nên trùng khớp hoàn toàn với tổng thể, nhưng thực tế thường bị thiếu sót (ví dụ danh bạ điện thoại bàn bỏ sót các hộ chỉ dùng điện thoại di động). **Điều tra toàn bộ** (census) thu thập dữ liệu từ *mọi* thành viên tổng thể; **điều tra chọn mẫu** (sample survey) chỉ thu thập từ một mẫu rồi suy rộng ra tổng thể. **Thống kê mẫu** (statistic) là con số tính được từ dữ liệu mẫu (như $\bar{x}$, $\hat{p}$); **tham số tổng thể** (parameter) là con số cố định nhưng thường không

biết, mô tả cả tổng thể (như μ, p). Mẹo nhớ: chữ cái đầu — *Statistic* đi với *Sample*, *Parameter* đi với *Population*.

Ví dụ mẫu · Population, sample, and sampling frame

A high school has $1{,}200$ enrolled students. The principal wants to estimate the average number of hours students spend on homework per week. She obtains a numbered roster of all $1{,}200$ student ID numbers from the registrar's office and uses a computer to randomly select 60 of those ID numbers; she then asks each of the 60 selected students to report their weekly homework hours, obtaining a sample mean of $\bar{x} = 8.4$ hours.

- **Population:** all $1{,}200$ enrolled students.
- **Sampling frame:** the registrar's numbered roster of $1{,}200$ student ID numbers (here it matches the population exactly, since every enrolled student has exactly one ID number).
- **Sample:** the 60 students who were selected and actually reported their hours.
- **Statistic:** $\bar{x} = 8.4$ hours (computed from the 60 sampled students).
- **Parameter being estimated:** μ, the true (unknown) mean weekly homework hours of all $1{,}200$ students.

This is a sample survey, not a census, because data was collected from only 60 of the $1{,}200$ students.

A census sounds like the most reliable option, since in principle no one is left out — but a true census is rarely practical. Contacting every single member of a large population is expensive and time-consuming, response rates are typically far from complete (so a "census" often becomes an incomplete sample survey in disguise), and by the time data collection finishes for a very large or fast-changing population, the population itself may already have changed. A carefully designed sample survey, using far fewer resources, can often estimate a population parameter with a documented and controllable amount of uncertainty — which is why sample surveys, not censuses, are the standard tool for most statistical questions.

3.2 Random Sampling Methods

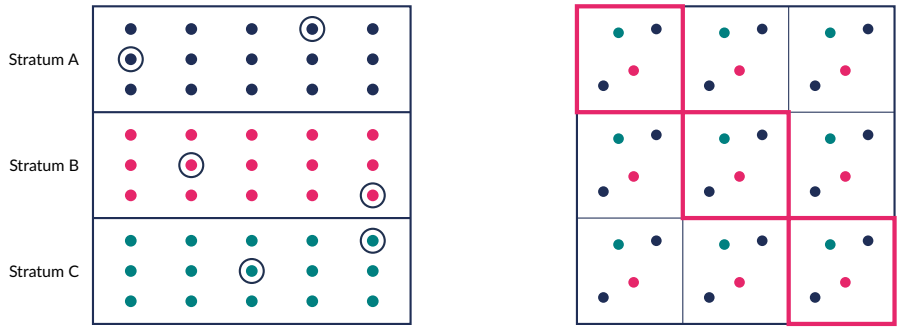
Random sampling is what allows a sample statistic to be used to draw a trustworthy conclusion about a population: because chance, not a person's judgment, decides who ends up in the sample, the sample tends to look like a smaller version of the population, and the size of the resulting error can be described mathematically. This is precisely why convenience samples — surveying whoever happens to be easiest to reach — are avoided in careful statistical work: without a chance mechanism behind the selection, there is no mathematical basis for saying how well the sample represents the population at all.

Simple Random Sample (SRS): every group of n individuals has an equal chance of being chosen. **Stratified random sample:** divide the population into homogeneous *strata*, then take an SRS from each. **Cluster sample:** divide into heterogeneous *clusters* that each represent the population, then randomly select whole clusters. **Systematic sample:** select every k-th individual from a list, starting at a random point.

VN **GIẢI THÍCH TIẾNG VIỆT**

Phân biệt các loại chọn mẫu: **SRS** — mọi nhóm n đối tượng có cơ hội bằng nhau; **phân tầng** — chia thành các tầng *giống nhau bên trong*, lấy SRS mỗi tầng; **cụm** — chia thành các cụm *đa*

dạng như cả tổng thể, chọn nguyên cụm. Sai lệch tự nguyện (voluntary response) luôn làm mẫu thiên về người có ý kiến **mạnh** (rất thích hoặc rất ghét).



Stratified sample:
strata are internally homogeneous

Cluster sample:
clusters each mirror the population

Left: a stratified sample divides the population into homogeneous strata (A, B, C) and draws an SRS from each. Right: a cluster sample divides the population into heterogeneous clusters that each resemble the whole population; entire clusters (bold outline) are then randomly selected in full.

3.2.1 Simple Random Sampling in Practice

Carrying out an SRS requires a numbered list of the population and a source of random digits (a random digit table or a random-number generator). Every member of the population is assigned a unique number using the same number of digits (e.g., $01, 02, \ldots, 28$ for a population of 28); digits are then read off in matching-length groups, discarding any group that is out of range or a repeat, until enough distinct individuals have been chosen.

Ví dụ mẫu · Selecting an SRS with random digits

A teacher has $N = 28$ students, numbered 01 through 28, and wants an SRS of $n = 5$ students to complete a short survey. She reads the following row from a random digit table, two digits at a time:

$$05913 \;\; 06227 \;\; 38841 \;\; 92107$$

Reading continuous two-digit groups from the left: $05,\ 91,\ 30,\ 62,\ 27,\ 38,\ 84,\ 19,\ 21,\ 07, \ldots$. Discard any group above 28 or equal to 00, and skip repeats:

- 05 — valid, select student #5.
- $91, 30, 62$ — all exceed 28, discard.
- 27 — valid, select student #27.
- $38, 84$ — exceed 28, discard.
- 19 — valid, select student #19.
- 21 — valid, select student #21.
- 07 — valid, select student #7.

Five distinct valid numbers have now been found, so the SRS is students #5, #7, #19, #21, #27. Because the digits were random and every valid group had an equal chance of appearing next, every student — and every possible group of 5 students — had an equal chance of being chosen.

3.2.2 Stratified Random Sampling in Practice

When a population naturally divides into subgroups that differ in a way related to the variable of interest, stratifying and sampling separately within each stratum guarantees every subgroup is represented and typically produces more precise estimates than an SRS of the same total size.

Ví dụ mẫu · Proportional stratified sampling

A district of $N = 1{,}000$ students consists of 600 urban students and 400 rural students; a researcher suspects internet access differs sharply between the two groups and wants an overall sample of $n = 50$, proportionally allocated. The number sampled from each stratum is proportional to its share of the population:

$$\text{Urban sample size} = 50 \times \frac{600}{1{,}000} = 30, \qquad \text{Rural sample size} = 50 \times \frac{400}{1{,}000} = 20.$$

The researcher numbers the 600 urban students 001–600 and draws an independent SRS of 30 of them using a random-number generator; separately, she numbers the 400 rural students 001–400 and draws an independent SRS of 20 of them. Combining the two SRSs gives the full stratified sample of $30 + 20 = 50$ students, with both urban and rural students guaranteed proportional representation.

3.2.3 Cluster Sampling in Practice

Cluster sampling is chosen not because it improves precision, but because it is often far cheaper or more practical: it avoids the need for a complete list of every individual in the population and concentrates data collection geographically.

Ví dụ mẫu · Selecting a cluster sample

A school district has 40 schools; district inspectors want to audit classroom textbook conditions but cannot afford to visit every classroom in every school. Each school's classrooms are similar in diversity to the district as a whole (a mix of grade levels, subjects, and building ages), so the schools can serve as clusters. Inspectors number the 40 schools and use a random-number generator to select 6 of them, then visit and inspect *every* classroom within those 6 selected schools. Each of the 40 schools had probability $\frac{6}{40} = \frac{3}{20} = 15\%$ of being one of the chosen clusters. Because whole schools — not individual classrooms scattered across the district — are visited, travel costs are dramatically reduced compared to a comparably sized SRS of classrooms drawn from all 40 schools.

3.2.4 Systematic Sampling in Practice

A systematic sample requires only that the population be arranged in some order (a list, a production line, a queue), a step size k, and one random starting point between 1 and k.

Ví dụ mẫu · Selecting a systematic sample

A factory produces $N = 5{,}000$ items per day, and quality control wants a systematic sample of $n = 50$ items. The step size is $k = N/n = 5{,}000/50 = 100$. A random-number generator picks a starting point r between 1 and 100; suppose $r = 37$. The sample then consists of items numbered $37, 137, 237, 337, \ldots, 4{,}937$ — every 100th item after the random start. Counting terms confirms the sample size: $\frac{4{,}937-37}{100} + 1 = 49 + 1 = 50$ items, as required.

A smaller-scale illustration: an ordered list of $N = 30$ items with step $k = N/n = 5$ and a random start $r = 3$. Items #3, #8, #13, #18, #23, #28 (highlighted) are selected, giving a sample of size $n = 6$.

(!) Lỗi thường gặp: Nếu danh sách tổng thể có một **quy luật lặp lại (periodicity)** trùng với bước nhảy k, mẫu hệ thống (systematic sample) có thể bị lệch nghiêm trọng dù quy trình chọn điểm bắt đầu là ngẫu nhiên. Ví dụ: một chuỗi nhà hàng ghi doanh thu mỗi ngày trong nhiều tuần; nếu lấy **every 7th day** (mỗi 7 ngày một lần), mẫu sẽ luôn rơi vào *cùng một thứ trong tuần* (ví dụ luôn là thứ Hai), khiến kết quả không đại diện cho các ngày còn lại — hoàn toàn không liên quan đến việc điểm bắt đầu có ngẫu nhiên hay không. Trước khi dùng mẫu hệ thống, cần kiểm tra xem thứ tự của danh sách có ẩn chứa chu kỳ nào trùng với k hay không.

3.2.5 Stratified versus Cluster: A Side-by-Side Comparison

Students frequently confuse these two designs because both begin by partitioning the population into groups. The difference is what those groups look like internally, and what gets randomly selected.

Feature	Stratified Random Sample	Cluster Sample
Groups formed	Homogeneous strata — individuals *within* a stratum are similar on a key variable	Heterogeneous clusters — each cluster is a small-scale mirror of the whole population
What is randomly selected	An SRS is drawn from *every* stratum	A random subset of *whole clusters*; every member of a chosen cluster is included
Main purpose	Guarantee representation of every subgroup and increase precision	Reduce cost/logistics when a full list of the population is unavailable or spread out
Typical use case	Sampling by grade level, income bracket, or region when subgroup comparisons matter	Sampling by school, city block, or store location when travel/access is expensive
Risk if poorly formed	Strata that are not truly homogeneous lose the precision benefit	Clusters that are not representative of the population bias the sample

VN **GIẢI THÍCH TIẾNG VIỆT**

Cả hai đều **chia nhóm** trước khi chọn, nhưng khác nhau ở bản chất nhóm và cách chọn: **phân tầng** chia thành tầng *đồng nhất bên trong* rồi lấy mẫu ngẫu nhiên *trong từng tầng* — mục tiêu là đảm bảo đại diện và tăng độ chính xác; **cụm** chia thành cụm *đa dạng như tổng thể* rồi chọn *nguyên cụm* — mục tiêu là tiết kiệm chi phí đi lại/khảo sát khi không có danh sách đầy đủ từng cá nhân.

3.2.6 Choosing Among the Four Methods

In practice, the choice of sampling method is driven by what is known about the population and what resources are available, not by any single "best" method. An **SRS** is the simplest to justify statistically and is the right default whenever a complete numbered list of the population is available and no important subgroup structure needs to be guaranteed. A **stratified sample** is preferred when the population contains subgroups that are known (or suspected) to differ substantially on the variable being measured, and comparing those subgroups — or ensuring every one of them is represented — matters to the study's purpose. A **cluster sample** is preferred when no complete list of individuals

exists, or when the population is spread out geographically and visiting individually selected people would be prohibitively expensive; the trade-off is some loss of precision compared to an SRS of the same size. A **systematic sample** is preferred when individuals naturally arrive or are already arranged in an accessible order (a checkout line, a production line, a sorted registry) and drawing a true SRS from that order would be impractical — provided the order contains no hidden periodicity that could align with the step size k.

3.3 Sources of Bias in Sampling

Even a study that begins with a genuinely random selection mechanism can still end up with a misleading sample if something goes wrong *after* selection — the sampling frame could already be incomplete, selected individuals could fail to respond, or respondents could answer inaccurately. Each of these failure points has a specific name.

Khái niệm cốt lõi

Bias is a systematic favoring of certain outcomes. **Undercoverage:** some groups are left out of the sampling frame. **Nonresponse bias:** chosen individuals cannot be reached or refuse to answer. **Response bias:** people answer untruthfully (e.g., due to wording, sensitive topics, interviewer presence). **Voluntary response bias:** the sample consists of people who choose themselves (e.g., call-in polls) — this tends to overrepresent strong opinions.

VN **GIẢI THÍCH TIẾNG VIỆT**

Phân biệt các loại sai lệch: **undercoverage** — một nhóm bị bỏ sót khỏi khung mẫu ngay từ đầu; **nonresponse** — người được chọn không trả lời được hoặc từ chối; **response bias** — người trả lời không trung thực (do cách hỏi, chủ đề nhạy cảm, có người phỏng vấn trực tiếp); **voluntary response** — mẫu gồm những người tự nguyện tham gia, luôn thiên về ý kiến **mạnh**.

Type of bias	Where the flaw occurs	Typical example
Undercoverage	In the sampling frame itself, before any selection happens	A phone survey using only landline numbers, missing cell-only households
Nonresponse bias	Selected individuals cannot be reached or refuse to participate	A mail survey with a very low return rate
Response bias	Respondents answer, but untruthfully or inaccurately	Leading question wording; sensitive topics; an interviewer's presence
Voluntary response bias	The sample is entirely self-selected by respondents	Online or call-in polls open to whoever chooses to respond

Wording effects are a particularly important cause of response bias: the exact phrasing of a question can push respondents toward a particular answer even when the underlying attitude has not changed.

Ví dụ mẫu · Leading versus neutral question wording

Neutral wording: "Do you support increasing funding for public parks?"
Leading (loaded) wording: "Given that public parks provide critical green space and important health benefits for our community, do you support increasing funding for public parks?"
The second version embeds a persuasive justification directly inside the question before asking for an opinion, pressuring respondents toward answering "yes" regardless of their actual views. Surveys using the leading version will tend to report a higher percentage in favor than surveys using the neutral version, even when sampled from the same population — this inflation is

response bias due to wording, not a genuine difference of opinion.

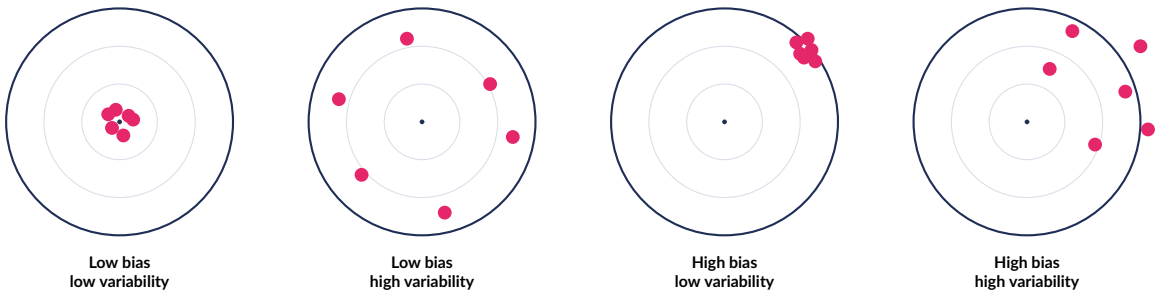
Ví dụ mẫu · Identify the bias

(a) A city government mails a paper survey about a proposed tax increase to all $5{,}000$ registered voters; only 400 are returned. *Diagnosis:* an 8% return rate means the vast majority of selected voters simply never responded — this is **nonresponse bias**, and the 400 who did respond may differ systematically (e.g., stronger opinions, more free time) from those who did not.
(b) A radio station invites listeners to call in and vote on whether the mayor is doing a good job. *Diagnosis:* the sample is entirely self-selected — only listeners motivated enough to call participate — so this is **voluntary response bias**, which typically overrepresents people with strong (often negative) opinions.
(c) A pollster estimates national opinion by calling only randomly generated *landline* telephone numbers. *Diagnosis:* younger adults, who disproportionately have only cell phones, are almost entirely excluded from the sampling frame before any calls are even made — this is **undercoverage**, a flaw in the frame itself rather than in who responds.
(d) A school nurse conducts face-to-face interviews with students, asking them directly whether they have ever vaped on school grounds. *Diagnosis:* the presence of an adult interviewer asking about a sensitive, rule-breaking behavior makes students more likely to underreport it — this is **response bias** due to the interview format itself (an anonymous, self-administered questionnaire on the same topic would likely produce more honest answers).

3.3.1 Bias versus Variability, and the Effect of Sample Size

Two very different questions can be asked about a sampling method: is it *centered* on the truth, and how much do its results *scatter* from sample to sample? These are the concepts of bias and variability, and confusing them is one of the most common errors in describing sampling methods.

Bias describes whether a sampling method tends to systematically overestimate or underestimate the population parameter — it is a property of the *method*, not of any single sample, and describes how far results are centered from the truth on average. **Variability** describes how much results from that method would differ from each other if the sampling were repeated many times — it describes the *spread* of results around wherever the method happens to be centered. **Increasing the sample size reduces variability** (repeated samples cluster more tightly together), **but it does not reduce bias**: a method with a structural flaw (e.g., voluntary response, undercoverage) remains just as systematically off no matter how large the sample gets, because the flaw is in *who* is included, not in *how many*.



Each target's center (small dark dot) is the true population parameter; each pink dot is the result of one repetition of a sampling method. Bias shifts the whole cluster of dots away from the center; variability spreads the dots out. Only reducing bias (choosing a better method) fixes off-center results — a larger sample size only tightens the spread.

Ví dụ mẫu · A famous case of a huge but biased sample

Before the 1936 U.S. presidential election, a magazine mailed roughly ten million straw-poll questionnaires to names drawn from telephone directories, magazine subscription lists, and car-registration records, and received about 2.4 million responses. Based on this enormous sample, it predicted a comfortable win for the Republican candidate; in reality, Franklin D. Roosevelt won in a landslide. The sample was not randomly selected — during the Depression, owning a telephone or a car skewed the frame toward wealthier, disproportionately Republican-leaning households (undercoverage) — and it also suffered from voluntary response bias, since only motivated recipients mailed their questionnaire back. No amount of additional responses could have fixed the prediction, because the flaw was bias in the sampling method, not insufficient sample size.

VN **GIẢI THÍCH TIẾNG VIỆT**

Bias (thiên lệch) là đặc điểm của *phương pháp* chọn mẫu — cho biết kết quả có bị lệch tâm khỏi giá trị thật hay không. **Variability** (độ biến thiên) cho biết kết quả dao động nhiều hay ít giữa các lần lặp lại. Tăng cỡ mẫu (n) chỉ làm **giảm độ biến thiên** (các kết quả tập trung gần nhau hơn), chứ **không sửa được bias** — nếu phương pháp chọn mẫu vốn đã thiên lệch (ví dụ voluntary response), mẫu càng lớn thì kết quả sai càng "chắc chắn" chứ không hề chính xác hơn. Đây là lý do một khảo sát với hàng triệu người trả lời tự nguyện vẫn có thể sai lệch nghiêm trọng hơn một mẫu ngẫu nhiên chỉ vài trăm người.

3.4 Observational Studies, Experiments, and the Language of Experiments

An **observational study** records data without imposing treatments and can only show *association*. An **experiment** deliberately imposes treatments and, if well designed, can support *causation*. The terms "explanatory variable" and "response variable" are used in *both* kinds of study — an observational study can still ask whether an explanatory variable is associated with a response variable — but only in an experiment does the researcher control who receives which value of the explanatory variable, which is what ultimately allows a causal conclusion.

The vocabulary of experiments

The **explanatory variable** (or independent variable) is the variable manipulated by the researcher, thought to explain or cause changes in another variable. The **response variable** (or dependent variable) is the outcome that is measured. A **treatment** is a specific condition applied to experimental units; when an experiment has more than one factor, a treatment is a specific *combination* of levels, one from each factor. An **experimental unit** is the individual object to which a treatment is applied; when the experimental units are people, they are called **subjects**. A **factor** is an explanatory variable that is deliberately manipulated in the experiment, and a **level** is one specific value, amount, or version of a factor. A **placebo** is a fake, inert treatment made to look identical to the real one, used to control for the **placebo effect** — the tendency of subjects to respond favorably to *any* treatment, real or fake, simply because they believe they are being treated.

VN **GIẢI THÍCH TIẾNG VIỆT**

Biến giải thích (explanatory variable) là biến bị tác động/thay đổi có chủ đích bởi người nghiên cứu; **biến đáp ứng** (response variable) là kết quả được đo lường. **Nghiệm thức** (treatment) là điều kiện cụ thể áp lên đối tượng thí nghiệm; nếu có nhiều **nhân tố** (factor), một nghiệm thức là một tổ hợp **mức độ** (level) của tất cả các nhân tố. **Đối tượng thí nghiệm** (experimental unit)

là vật/cá thể nhận nghiệm thức; nếu là người thì gọi là **subject**. **Giả dược** (placebo) là nghiệm thức "giả" trông giống thật, dùng để kiểm soát **hiệu ứng giả dược** (placebo effect) — xu hướng người tham gia cảm thấy đỡ hơn chỉ vì tin rằng mình đang được điều trị, dù thực tế không nhận thuốc thật.

Ví dụ mẫu · Labeling the language of an experiment

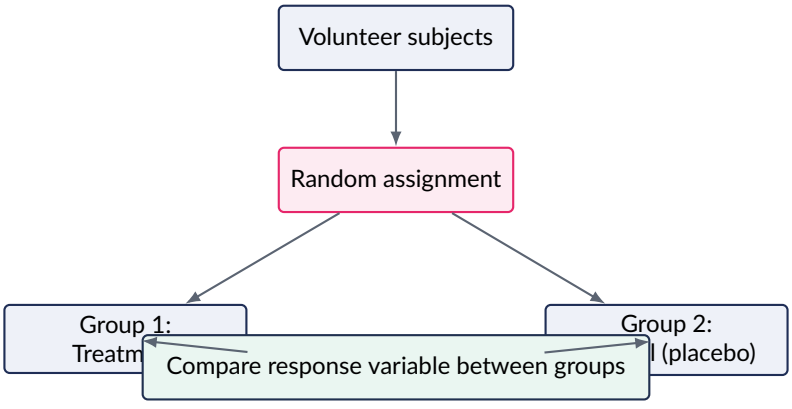
A researcher tests whether a sleep-tracking app improves sleep quality. She recruits 60 adult volunteers; 30 are randomly assigned to use the app nightly for 4 weeks, and the other 30 keep a simple sleep diary with no app. At the end of 4 weeks, every participant rates their sleep quality on a 0–100 scale.

- **Explanatory variable (factor):** whether or not the sleep app is used.
- **Levels of the factor:** "app" and "no app" — 2 levels.
- **Treatments:** since there is only one factor, the treatments are simply the 2 levels: "app" and "no app" (control).
- **Response variable:** the self-reported sleep quality score (0–100).
- **Experimental units / subjects:** the 60 adult volunteers.
- **Placebo:** not really possible here — subjects obviously know whether they are using an app, so this design cannot be blinded on the subject side (a genuine limitation, discussed further with blinding below).

When an experiment has *more than one* factor, a treatment is a specific combination of levels, one from each factor. For example, if the factors are fertilizer amount (levels: low, high) and watering frequency (levels: daily, weekly), there are $2 \times 2 = 4$ treatments in all: low-daily, low-weekly, high-daily, high-weekly.

3.5 Principles of Experimental Design

Principles of good experimental design: comparison (a control group), **random assignment** of subjects to treatments, and **replication** (enough subjects per group). A **confounding variable** affects the response and is related to the explanatory variable, making it impossible to separate their effects. **Blocking** groups subjects who are similar on a variable expected to affect the response, then randomizes *within* each block. A **double-blind** design keeps both subjects and evaluators unaware of treatment assignment, controlling for the placebo effect and evaluator bias.



A completely randomized experimental design with a placebo control group.

VN **GIẢI THÍCH TIẾNG VIỆT**

Bốn nguyên tắc thiết kế thí nghiệm tốt: **so sánh** (comparison) — luôn có nhóm đối chứng (control); **ngẫu nhiên hoá** (random assignment) — dùng cơ chế ngẫu nhiên để chia đối tượng vào các nhóm nghiệm thức, tránh thiên vị của người thực hiện; **lặp lại** (replication) — đủ số đối tượng mỗi nhóm để sai số ngẫu nhiên "trung hoà" lẫn nhau; và **kiểm soát các biến ngoại lai** — giữ mọi điều kiện khác giống nhau giữa các nhóm ngoại trừ nghiệm thức đang thử nghiệm, để tránh **nhiễu biến** (confounding variable).

Ví dụ mẫu · Designing a completely randomized experiment

A teacher wants to test whether listening to instrumental background music improves performance on a timed math worksheet, using 50 available student volunteers.

- **Treatments:** (1) complete the worksheet while instrumental music plays; (2) complete the worksheet in silence (control).
- **Random assignment method:** number the 50 volunteers 01–50; using a random-number generator, draw 25 distinct numbers without replacement — those students form the music group, and the remaining 25 form the silence group.
- **Response variable:** the number of correctly answered problems on the worksheet.
- **Replication:** using 25 students per group (rather than just 1) allows ordinary variation in individual math ability to average out, so any consistent difference between groups is more likely due to the music itself.
- **Control of extraneous variables:** both groups receive the identical worksheet, the identical time limit, and the identical room conditions except for the presence or absence of music, so any confounding from unequal conditions is avoided.

Because the students themselves know whether they are hearing music, this design cannot be blinded on the subject side; however, whoever grades the worksheets *can* be kept unaware of which group produced which paper, which is discussed further below.

(!) Lỗi thường gặp: **Replication** (nhiều đối tượng trên mỗi nhóm nghiệm thức, trong *cùng một* thí nghiệm) hoàn toàn khác với **replicating a study** (lặp lại toàn bộ nghiên cứu ở một thời điểm/địa điểm/nhóm đối tượng khác, thường là một nghiên cứu *độc lập* sau này). Một thí nghiệm với 25 người mỗi nhóm đã có **replication** — không cần thêm một thí nghiệm hoàn toàn mới để thoả điều kiện này. Đề thi hay đánh lừa bằng cách hỏi "cần làm gì để tăng replication?" — câu trả lời đúng là **tăng số đối tượng thí nghiệm trong mỗi nhóm**, không phải "lặp lại thí nghiệm ở trường khác."

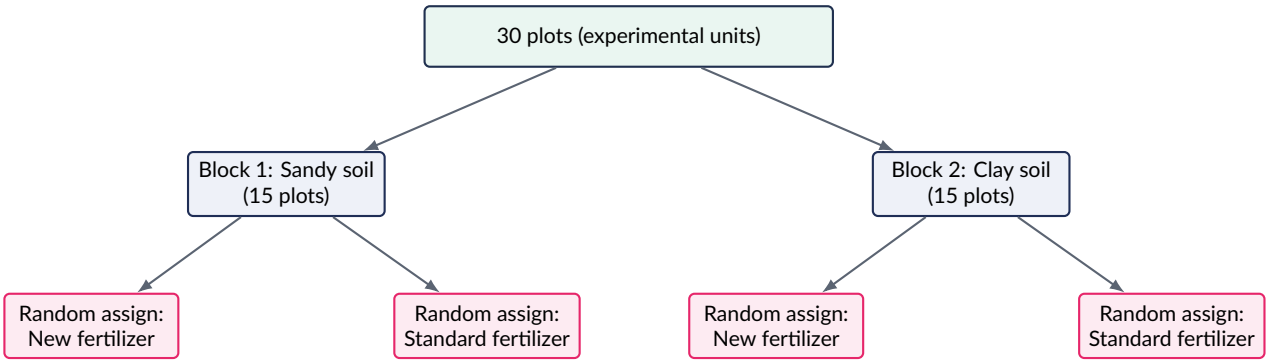
3.6 Randomized Block Design and Matched-Pairs Design

3.6.1 Randomized Block Design

Sometimes researchers already know, before the experiment even begins, that some other variable — not the one being tested — will strongly affect the response. **Blocking** groups experimental units into blocks that are similar with respect to that nuisance variable *before* treatments are assigned, and randomization is then carried out *separately within each block*. This is the experimental analogue of stratified sampling: it removes a known source of variability from the treatment comparison, increasing the experiment's precision.

Ví dụ mẫu · Extending the fertilizer experiment with blocking

Researchers test a new fertilizer against the standard fertilizer on 30 plots of land, but 15 of the plots have sandy soil and 15 have clay soil — a factor known to strongly affect plant growth on its own. Rather than randomly assigning all 30 plots at once, they first **block** by soil type: within the 15 sandy plots, they randomly assign about half to the new fertilizer and the rest to the standard fertilizer; independently, within the 15 clay plots, they again randomly assign about half to each fertilizer. Growth is compared between the two fertilizers *within* the sandy block and *within* the clay block separately, and the two within-block comparisons are then combined. Because soil-type variability is now balanced out within each block rather than contaminating the overall comparison, any remaining difference in growth is more clearly attributable to the fertilizer itself.



A randomized block design: plots are first sorted into blocks by soil type (the nuisance variable), and treatments are randomly assigned separately *within* each block, isolating the fertilizer's effect from soil-type variability.

3.6.2 Matched-Pairs Design

A **matched-pairs design** is a special case of blocking in which every block has exactly two experimental units. There are two common versions: (1) the same subject is measured under *both* treatments, often in a randomly determined order to avoid an order effect; or (2) two very similar subjects — matched on characteristics expected to affect the response — are paired together, and within each pair, one is randomly assigned to each treatment.

> In a matched-pairs design, "random assignment" still happens — it just happens *within each pair* rather than across the whole group. For a same-subject design, this typically means randomly choosing which treatment the subject receives first (or which side/location on the subject receives which treatment), so that order or position does not become a confounding variable.

For a same-subject, before/after style matched pair, order matters: if every subject always received Treatment A first and Treatment B second, any change over time (fatigue, learning, or simply the passage of time) would be confounded with the treatment difference. Suppose 20 volunteers each test two brands of noise-cancelling headphones for perceived noise reduction. For each volunteer, a coin flip decides whether Brand A or Brand B is tested first; the two within-person ratings are then compared. Randomizing the order — rather than testing every volunteer on Brand A first — prevents an order effect from being confused with a genuine brand difference.

> **(!) Lỗi thường gặp:** Không phải cứ dùng **khối** (block) là luôn tốt hơn. Nếu biến dùng để chia khối thực ra **không** ảnh hưởng đáng kể đến biến đáp ứng, việc chia khối không làm sai kết quả, nhưng cũng không mang lại lợi ích gì — chỉ làm thiết kế phức tạp hơn một cách không cần thiết so với thiết kế hoàn toàn ngẫu nhiên (completely randomized design) đơn giản.

3.6.3 Block Design or Completely Randomized? A Worked Comparison

Ví dụ mẫu · Choosing between a block design and a completely randomized design

(a) A dermatologist wants to compare a new anti-itch cream to the standard cream using 24 volunteers, each of whom has a similar itchy rash on *both* forearms. For each volunteer, a coin flip decides which forearm receives the new cream and which receives the standard cream; after two weeks, a blinded evaluator (who does not know which arm received which cream) rates the itchiness of each arm. *Diagnosis:* this is a **matched-pairs design** — each volunteer serves as their own block of size 2, controlling for person-to-person differences in skin sensitivity, and the coin flip is the random assignment step *within* each pair. Randomizing which arm gets which cream (rather than always using, say, the left arm for the new cream) prevents any systematic left/right difference from confounding the comparison.
(b) An online retailer tests 3 website layouts (A, B, C) on conversion rate using 300 shoppers. The retailer suspects returning customers react very differently to layout changes than new customers, so it first separates the 300 shoppers into "returning" and "new" groups, then randomly assigns shoppers to a layout *separately within each group*. *Diagnosis:* this is a **randomized block design** with customer type as the blocking variable — blocking is used (rather than a simple completely randomized design across all 300 shoppers at once) because customer type is expected to strongly affect the response, so isolating its effect increases the precision of the layout comparison.
(c) If the retailer had ignored the returning/new distinction entirely and simply randomly assigned all 300 shoppers to one of the 3 layouts, that would instead be a **completely randomized design** — still valid, but without the extra precision gained from blocking on a variable known to affect the response.

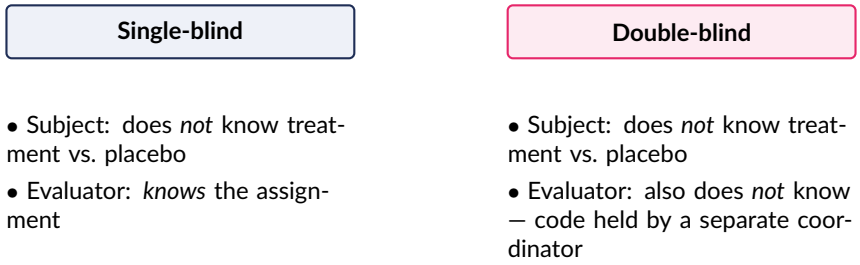
VN **GIẢI THÍCH TIẾNG VIỆT**

Thiết kế khối (block design) chia đối tượng thành các khối *giống nhau bên trong* theo một biến gây nhiễu đã biết trước, rồi mới ngẫu nhiên hoá *trong từng khối* — giống hệt logic của chọn mẫu phân tầng nhưng áp dụng cho việc *phân nhóm nghiệm thức*. **Thiết kế cặp ghép** (matched pairs) là trường hợp đặc biệt của thiết kế khối khi mỗi khối chỉ có 2 đối tượng — có thể là cùng một người đo hai lần (trước/sau), hoặc hai người rất giống nhau ghép cặp lại. Chỉ nên dùng khối khi biến khối thực sự ảnh hưởng đến biến đáp ứng; nếu không, thiết kế hoàn toàn ngẫu nhiên (completely randomized design) đơn giản hơn mà vẫn hợp lệ.

3.7 Confounding Variables, Blinding, and the Placebo Effect

A **confounding variable** is a variable — often one the researcher never measured — that affects the response variable *and* is related to the explanatory variable, so that its effect on the response cannot be separated from the effect of the explanatory variable. Confounding is the single biggest threat to drawing causal conclusions from an observational study. It is a distinct problem from the sampling biases of Section 3: bias in sampling is a flaw in *who ends up in the data*, while confounding is a flaw in *the comparison itself* — a third variable, unrelated to how the data were collected, is tangled up with the explanatory variable being studied. Random assignment is the tool that specifically defends against confounding, because it tends to spread any lurking variable roughly evenly across treatment groups.

Single-blind design: only *one* side (typically the subjects) is kept unaware of the treatment assignment. **Double-blind** design: *neither* the subjects *nor* the people evaluating outcomes know who received which treatment; the assignment code is usually held by a separate, uninvolved study coordinator. Double-blinding controls for two distinct problems at once: the **placebo effect** (subjects who believe they are being treated tend to report improvement even from an inert treatment) and **evaluator bias** (a researcher who knows a subject received the "real" treatment may — even unconsciously — score their outcome more favorably).



Single- versus double-blind designs. Double-blinding controls for both the placebo effect (subject-side bias) and evaluator bias (unconscious favoritism when scoring or diagnosing outcomes).

Ví dụ mẫu · Spot the confounding variable

A news report states: "Students who eat breakfast every day have higher GPAs than students who skip breakfast, so schools should require students to eat breakfast to raise grades." This was purely an observational study — no one randomly assigned students to eat breakfast or skip it.
Possible confounding variable: household routine and socioeconomic stability. Families with structured mornings, consistent schedules, and reliable access to food are more likely both to ensure their children eat breakfast *and* to provide other advantages (sleep, tutoring, stable housing) that independently raise GPA. Because breakfast-eating was not randomly assigned, this lurking variable could account for some or all of the observed association — so the claim that eating breakfast *causes* higher GPA is not justified by this study alone. To turn this into a study that *could* support a causal claim, researchers would need to randomly assign otherwise-similar students to either eat breakfast or skip it — precisely the random-assignment step that this observational study never performed.

3.8 Scope of Inference

Scope of inference: random **sampling** allows generalizing results to the population; random **assignment** allows concluding cause-and-effect. An observational study — even a large, well-conducted one — can never establish causation because of possible confounding.

(!) Lỗi thường gặp: "Random sample" và "random assignment" là hai việc **khác nhau**: lấy mẫu ngẫu nhiên cho phép suy rộng ra tổng thể; chỉ có phân nhóm ngẫu nhiên (random assignment) trong thí nghiệm mới cho phép kết luận **nhân quả**. Có cả hai mới vừa suy rộng vừa kết luận nhân quả được.

The two decisions — whether individuals were randomly *sampled* from a population, and whether they were randomly *assigned* to treatments — are completely independent of each other, giving four possible combinations and four correspondingly different sets of valid conclusions.

	Random assignment: Yes	Random assignment: No
Random sampling: Yes	Can generalize to the population *and* infer causation (the strongest possible design, though rare in practice)	Can generalize to the population; *cannot* infer causation (a random-sample survey or observational study)
Random sampling: No	Can infer causation *for the subjects studied*; cannot safely generalize to a larger population (a typical volunteer-based experiment)	Cannot generalize *and* cannot infer causation (an observational study on a convenience sample)

Ví dụ mẫu · One scenario for each of the four cases

(a) Random sampling: yes; random assignment: yes. Researchers randomly select 200 hospitals nationwide, then within each hospital randomly assign patients undergoing a certain surgery to Drug A or Drug B. *Conclusion:* the results can be generalized to patients at hospitals nationwide, and the difference in recovery outcomes can be attributed to (caused by) the drug.
(b) Random sampling: yes; random assignment: no. A survey randomly samples $1{,}000$ U.S. adults and simply asks how many hours they exercise weekly and how they rate their stress level; no treatment is imposed. *Conclusion:* the association between exercise and stress can be generalized to U.S. adults, but exercise cannot be said to *cause* lower stress, since this is purely observational.
(c) Random sampling: no; random assignment: yes. A professor recruits 40 students from her own class (not randomly selected from all students) and randomly assigns 20 to a new memory technique and 20 to normal studying. *Conclusion:* for these 40 students specifically, the technique appears to have caused the improvement in memory performance, but the result cannot be safely generalized to all students, since the 40 were not randomly sampled from a larger population.
(d) Random sampling: no; random assignment: no. A blogger reads through comments voluntarily left by readers describing their diet and mood. *Conclusion:* neither generalization (the readers were self-selected, not randomly sampled) nor causation (no treatment was randomly assigned) is justified — the weakest possible design.

VN **GIẢI THÍCH TIẾNG VIỆT**

Bốn ô trong bảng 2×2 trên tương ứng với bốn mức độ suy luận khác nhau: có **cả hai** (lấy mẫu ngẫu nhiên + phân nhóm ngẫu nhiên) thì vừa suy rộng ra tổng thể vừa kết luận được nhân quả; chỉ có **lấy mẫu ngẫu nhiên** (không có phân nhóm) thì suy rộng được nhưng không kết luận nhân quả; chỉ có **phân nhóm ngẫu nhiên** (mẫu không ngẫu nhiên) thì kết luận nhân quả được nhưng chỉ đúng cho nhóm đối tượng đã nghiên cứu, không suy rộng được; và nếu **không có gì cả** thì không được kết luận điều gì chắc chắn ngoài những gì quan sát trực tiếp trên chính mẫu đó.

3.9 Ethical Considerations in Data Collection

Any study involving human participants carries ethical obligations that are independent of its statistical design. Participants must give **informed consent** — they must be told the purpose, procedures, and any foreseeable risks of the study and agree to take part before data collection begins. Participants must retain the **right to withdraw** from the study at any time without penalty. All individual data must be kept **confidential**, with results reported only in aggregate or anonymized form. For studies conducted at universities, hospitals, and most research organizations, the proposed design must first be approved by an **Institutional Review Board (IRB)** — an independent panel that reviews

the procedures in advance to confirm that risks to participants are minimized and are reasonably justified by the value of the knowledge the study may produce. Additional safeguards apply to vulnerable populations, such as children or hospitalized patients, who typically require consent from a parent or guardian in addition to the participant's own agreement; if any element of deception is used in a study's design, participants are ordinarily debriefed — told the true purpose of the study — as soon as their participation ends.

3.10 Practice Problems

Bài tập luyện tập

A school wants to survey student opinions and randomly selects 5 whole homerooms out of 40, then surveys every student in each selected homeroom. This is:

(A) A simple random sample
(B) A stratified random sample
(C) A cluster sample
(D) A systematic sample

Lời giải chi tiết

Homerooms are heterogeneous groups that each roughly reflect the student body, and *entire* homerooms are randomly selected (not individuals within each). This is a **cluster sample**. **(C)**.

Bài tập luyện tập

A university registrar assigns each of the $8{,}000$ enrolled students a unique ID number and uses a computer's random-number generator to select 200 ID numbers, surveying those 200 students about campus dining. This sampling method is:

(A) A simple random sample
(B) A stratified random sample
(C) A cluster sample
(D) A systematic sample

Lời giải chi tiết

Every student has an equal chance of selection, and every possible group of 200 students has an equal chance of being the chosen group — this is the defining property of an SRS. **(A)**.

Bài tập luyện tập

A hospital administrator knows patient satisfaction differs a great deal between the pediatric, adult, and geriatric wards. She separately draws an SRS of 15 patients from each of the three wards. This sampling method is:

(A) A simple random sample
(B) A stratified random sample
(C) A cluster sample
(D) A systematic sample

Lời giải chi tiết

The wards form homogeneous groups (patients within a ward tend to be similar), and an SRS is drawn separately from *every* group — the defining property of a stratified random sample. **(B)**.

Bài tập luyện tập

A quality inspector examines every 25th car coming off an assembly line, beginning with the 8th car. This sampling method is:

(A) A cluster sample
(B) A stratified random sample
(C) A simple random sample
(D) A systematic sample

Lời giải chi tiết

Individuals are chosen at a fixed interval ($k = 25$) from an ordered list, starting at a randomly chosen point (the 8th car) — this is a systematic sample. **(D)**.

Bài tập luyện tập

A researcher has a complete list of all 120 neighborhoods in a large county, each of which contains a socioeconomically diverse mix of residents similar to the county as a whole. She randomly selects 10 neighborhoods and surveys every household within them, rather than trying to obtain a list of all residents individually. Explain why this is a cluster sample rather than a stratified sample.

Lời giải chi tiết

The 120 neighborhoods are **heterogeneous** — each one internally mirrors the diversity of the whole county — rather than homogeneous, so they function as clusters, not strata. Furthermore, *whole* neighborhoods are randomly selected and every household within a chosen neighborhood is included, rather than an SRS being drawn from within every neighborhood (which is what stratification would require). Both features together — heterogeneous groups, and selecting entire groups rather than sampling within each one — identify this as a cluster sample.

Bài tập luyện tập

A researcher mails a survey to 2000 randomly selected households; only 300 return it, and the topic (household debt) may make respondents unwilling to answer honestly. The two biases most clearly present are:

(A) Undercoverage and voluntary response
(B) Nonresponse bias and response bias
(C) Response bias and undercoverage
(D) Voluntary response and systematic bias

Lời giải chi tiết

Only a small fraction returned the survey — **nonresponse bias** — and a sensitive topic like debt may cause untruthful answers — **response bias**. **(B)**.

Bài tập luyện tập

A news website posts an online poll asking "Do you support the new city budget?" and reports results based on whoever chooses to click a response. This is an example of:

(A) Nonresponse bias
(B) Undercoverage
(C) Voluntary response bias
(D) Systematic sampling error

Lời giải chi tiết

The sample consists entirely of people who chose to participate on their own; such self-selected samples overrepresent people with strong opinions. **(C)**.

Bài tập luyện tập

A political pollster surveys opinion by calling only randomly generated *landline* telephone numbers. Younger adults, who disproportionately have only cell phones, are almost entirely absent from the sample. This is an example of:

(A) Voluntary response bias
(B) Response bias
(C) Undercoverage
(D) Nonresponse bias

Lời giải chi tiết

The flaw is in the sampling frame itself (landline numbers) before any calls are even placed — an entire group is structurally excluded from ever being selected. **(C)**.

Bài tập luyện tập

A city council considers two versions of a survey question: *Version 1:* "Do you support building a new public library?" *Version 2:* "Given that literacy programs benefit children throughout our community, do you support building a new public library?" Explain what type of bias Version 2 is likely to introduce, and why.

Lời giải chi tiết

Version 2 is a **leading (loaded) question**, which produces **response bias** due to wording. By embedding a persuasive justification directly into the question before asking for an opinion, it pressures respondents toward answering "yes," inflating the apparent level of support compared with the neutrally worded Version 1 — even though the underlying opinions of the population have not changed.

Bài tập luyện tập

A polling company wants to estimate the proportion of all 5.2 million registered voters in a state who favor a ballot measure. It surveys a random sample of 800 voters and finds that 62% favor the measure. The value 62% is:

(A) A population
(B) A parameter
(C) A statistic
(D) A census

Lời giải chi tiết

The 62% was computed from the 800 sampled voters, not from all 5.2 million registered voters — any number computed from sample data is a statistic (here, it estimates the unknown population parameter, the true proportion of all 5.2 million voters who favor the measure). **(C)**.

Bài tập luyện tập

The United States government attempts to count and collect basic data from every single resident of the country every ten years. This is an example of a:

(A) A sample survey
(B) A census
(C) A cluster sample
(D) A voluntary response sample

Lời giải chi tiết

Data is collected from (attempted for) *every* member of the population, not just a subset — this is a census by definition. **(B)**.

Bài tập luyện tập

Students who drink more coffee tend to have higher stress scores. A researcher wants to claim coffee *causes* stress. What is the strongest objection?

(A) Correlation was not computed
(B) This is an observational study, so a confounding variable (e.g., workload) could explain both
(C) The sample size is too small
(D) The response variable is categorical

Lời giải chi tiết

Without random assignment of a "treatment" (coffee amount), this is an observational study; a lurking/confounding variable such as workload or sleep could independently raise both coffee intake and stress. Association $\neq$ causation. **(B)**.

Bài tập luyện tập

Neighborhoods with more ice cream shops per capita also tend to have higher rates of swimming-pool drownings. A newspaper op-ed claims ice cream shops cause drownings and calls for tighter permits. Which lurking variable most plausibly confounds this relationship?

(A) Neighborhood population density
(B) Outdoor temperature / season (hot weather drives both ice cream sales and swimming)
(C) The price of ice cream
(D) The number of lifeguards on duty

Lời giải chi tiết

Hot weather independently increases both ice cream purchases and swimming (and therefore drowning risk); it is related to the "explanatory" variable (ice cream sales) and directly affects the response (drownings), the exact definition of a confounding variable. **(B)**.

Bài tập luyện tập

In a clinical trial for a new pain medication, neither the patients nor the doctors evaluating pain levels know who received the actual drug versus an inactive pill that looks identical. This design primarily protects against:

(A) Undercoverage in the sampling frame
(B) The placebo effect and evaluator bias
(C) Nonresponse bias
(D) Voluntary response bias

Lời giải chi tiết

Keeping both subjects and evaluators unaware of assignment is the definition of double-blinding, which controls simultaneously for the placebo effect (subject-side) and evaluator bias (evaluator-side). **(B)**.

Bài tập luyện tập

A wildlife researcher records the diet and body weight of wild bears exactly as she finds them in the forest, without altering their environment or diet in any way. This is:

(A) A completely randomized experiment
(B) An observational study
(C) A matched-pairs experiment
(D) A randomized block design

Lời giải chi tiết

No treatment is imposed on the bears; data is simply recorded as it naturally occurs. Without an imposed treatment, this cannot be an experiment of any kind. **(B)**.

Bài tập luyện tập

A food scientist tests how oven temperature (325°F or 375°F) and baking time (20 or 25 minutes) affect the moistness of muffins. Batches of muffin batter are randomly assigned to one of the four temperature-time combinations. Identify: (a) the factors, (b) the number of levels of each factor, (c) the number of treatments, (d) the experimental units, (e) the response variable.

Lời giải chi tiết

(a) The factors are oven temperature and baking time. (b) Each factor has 2 levels (325°F/375°F; 20 min/25 min). (c) A treatment is a combination of one level from each factor, so there are $2 \times 2 = 4$ treatments: 325°F-20min, 325°F-25min, 375°F-20min, 375°F-25min. (d) The experimental units are the batches of muffin batter. (e) The response variable is moistness (however it is measured, e.g., a moisture-content score).

Bài tập luyện tập

Explain the difference between an "experimental unit" and a "subject," using an experiment that applies fertilizer to individual plant pots and an experiment that gives a new memory drug to individual people.

Lời giải chi tiết

"Experimental unit" is the general term for whatever a treatment is applied to — it could be a person, an animal, a plot of land, or a batch of material. "Subject" refers specifically to a *human* experimental unit. In the fertilizer experiment, the experimental units are the plant pots — they are not called subjects, since they are not human. In the memory-drug experiment, the experimental units are also correctly called subjects, because the individuals receiving the drug

are people.

Bài tập luyện tập

Researchers test a new fertilizer on plant growth, but suspect that soil type (sandy vs. clay) also strongly affects growth. Describe an experimental design that accounts for this using blocking.

Lời giải chi tiết

Block the plots by soil type (sandy, clay) since soil type is expected to affect the response. **Within each block**, randomly assign plots to receive the new fertilizer or the standard fertilizer (control). Compare growth between treatment and control *within* each soil block, then combine results. This randomized block design removes soil-type variability from the treatment comparison, increasing the experiment's precision.

Bài tập luyện tập

A dermatologist wants to compare a new anti-itch cream (Cream X) to the standard cream (Cream Y), using 24 volunteers who each have a similar itchy rash on both forearms. Design a matched-pairs experiment.

Lời giải chi tiết

For each of the 24 volunteers, flip a coin to randomly decide which forearm receives Cream X and which receives Cream Y — each volunteer is their own matched pair (a block of size 2), which controls for person-to-person differences in skin sensitivity and rash severity. After a fixed treatment period, a blinded evaluator who does not know which arm received which cream rates the itchiness of each arm using a standardized scale. The comparison is then made *within* each volunteer (the difference between their two arm scores), and these 24 within-person differences are combined. Randomizing which arm receives which cream (rather than a fixed rule, such as always treating the left arm with Cream X) prevents any systematic left/right effect from confounding the comparison.

Bài tập luyện tập

A teacher wants to test whether listening to instrumental background music improves performance on a timed math worksheet, using 50 available student volunteers. Design a completely randomized experiment, clearly identifying the treatments, the method of random assignment, and the response variable.

Lời giải chi tiết

Treatments: (1) complete the worksheet while instrumental music plays; (2) complete the worksheet in silence (control). **Random assignment:** number the 50 volunteers 01–50; use a random-number generator to select 25 distinct numbers without repeats to form the music group, with the remaining 25 students forming the silence group. **Response variable:** the number of correctly answered problems on the worksheet. All 50 students should receive the identical worksheet, time limit, and room conditions except for the music/silence manipulation, so that any consistent difference in scores between the two groups can be attributed to the music rather than to some other varying condition.

Bài tập luyện tập

A physical-education teacher wants to compare two warm-up routines' effect on 100-meter sprint times, using 40 students — 20 of whom are on the school's competitive track team (much faster on average) and 20 of whom are not. Should she use a completely randomized design or a randomized block design? Justify your choice and describe how you would carry it out.

Lời giải chi tiết

She should use a **randomized block design**, blocking by track-team status. Track-team membership is a nuisance variable already known to strongly affect sprint times, independent of the warm-up routine, so leaving it uncontrolled would add unnecessary variability to the treatment comparison. **Procedure:** form two blocks — "track team" (20 students) and "non-track-team" (20 students). Within the track-team block, randomly assign 10 students to Warm-up A and 10 to Warm-up B; independently, within the non-track-team block, randomly assign 10 students to Warm-up A and 10 to Warm-up B. Compare sprint times between the two warm-ups *within* each block, then combine the two within-block comparisons. This isolates the effect of the warm-up routine from the large speed difference that track-team status alone would otherwise create.

Bài tập luyện tập

Before a university study involving human participants can begin, the researchers must submit their proposed procedures for review by an independent panel that evaluates whether risks to participants are justified and reasonably minimized. This panel is called:

(A) A sampling frame committee
(B) An Institutional Review Board (IRB)
(C) A cluster audit committee
(D) A census bureau

Lời giải chi tiết

This independent, prior-approval review of a study's ethics and risk-to-benefit balance is exactly the role of an Institutional Review Board. **(B)**.

Bài tập luyện tập

A polling organization takes a simple random sample of 1500 registered voters nationwide and asks whether they approve of a proposed law; it does not assign anyone to any treatment. Based on this study, which conclusion is justified?

(A) Can generalize to all registered voters, and the wording of the law caused their opinions
(B) Can generalize to all registered voters, but cannot establish that any single factor caused their opinions
(C) Cannot generalize beyond the 1500 voters, but causation can be determined
(D) Cannot generalize and cannot determine causation

Lời giải chi tiết

Random **sampling** was used (so results generalize to all registered voters), but there was no random **assignment** of any treatment (this is purely observational), so no cause-and-effect

conclusion is justified. **(B)**.

Bài tập luyện tập

A biology professor recruits (via a voluntary sign-up sheet, not a random sample of all students) 60 students from her own university to test a new tutoring method, then randomly assigns 30 to receive the tutoring and 30 to a control group. Test scores are significantly higher in the tutoring group. What conclusions are justified, and what conclusion is *not* justified?

Lời giải chi tiết

Random **assignment** was used (30 vs. 30, randomly split), so for these particular 60 students, the tutoring method can be said to have **caused** the improvement in test scores. However, random **sampling** was *not* used — the 60 students were self-selected volunteers, not randomly drawn from a larger population of students — so the result cannot be safely generalized to all students in general; it applies only to students similar to those who volunteered.

Bài tập luyện tập

Researchers obtain a list of every hospital in a state and randomly select 40 of them. Within each selected hospital, patients undergoing a certain surgery are randomly assigned to receive either a new post-operative pain protocol or the standard protocol, and recovery time is compared. Classify this study using the random sampling / random assignment framework, and state precisely what conclusions are justified.

Lời giải chi tiết

Random sampling: yes (40 hospitals randomly selected from all hospitals in the state). **Random assignment: yes** (patients randomly assigned to a protocol within each hospital). This is the top-left cell of the framework — the strongest possible position. **Conclusions:** the results can be generalized to (patients undergoing this surgery at) hospitals throughout the state, and because of random assignment, any consistent difference in recovery time can be attributed to (caused by) the pain protocol rather than to some other lurking variable.

Bài tập luyện tập

A club has 24 members, and an SRS of 4 members will be selected to form a committee. What is the probability that two particular members, Alice and Bob, are *both* selected for the committee?

Lời giải chi tiết

The total number of possible committees of 4 from 24 members is $\binom{24}{4} = 10{,}626$. The number of committees that include *both* Alice and Bob is the number of ways to choose the remaining 2 members from the other 22: $\binom{22}{2} = 231$. So

$$P(\text{both Alice and Bob selected}) = \frac{231}{10{,}626} = \frac{1}{46} \approx 0.0217.$$

Check by a second method: $P(\text{Alice selected}) = \frac{4}{24} = \frac{1}{6}$; given Alice is selected, 3 spots remain among the other 23 members, so $P(\text{Bob also selected} \mid \text{Alice selected}) = \frac{3}{23}$. Multiplying, $P(\text{both}) = \frac{1}{6} \times \frac{3}{23} = \frac{3}{138} = \frac{1}{46}$, confirming the answer.

Bài tập luyện tập

A grocery chain records total daily sales for one of its stores and wants a systematic sample of 8 days from a 56-day (8-week) stretch to estimate typical daily sales. An employee proposes using $k = 7$ (one week) as the step size, with a randomly chosen start day in the first week. Explain why this specific choice of k is a poor idea, even though the starting day is chosen randomly.

Lời giải chi tiết

Because a week has exactly 7 days, a step size of $k = 7$ means every single day selected will fall on the **same day of the week** (e.g., if the random start is a Saturday, every selected day is a Saturday). Daily sales are known to follow a weekly pattern (weekends typically differ from weekdays), so this sample would badly misrepresent the store's overall daily sales, systematically over- or under-estimating the true average depending on which day of the week happened to be selected. Randomizing the starting point does not fix this problem, because the flaw is that the population list (daily sales) has a periodicity that matches k — a different step size (e.g., $k = 8$ or $k = 9$, which does not divide evenly into a 7-day cycle) would avoid always landing on the same weekday.

Bài tập luyện tập

Two sampling methods are used repeatedly to estimate the same unknown population proportion p. Method A's results are tightly clustered but consistently centered well above the true value of p. Method B's results are centered right on the true value of p, but scattered widely from sample to sample. (a) Describe Method A and Method B in terms of bias and variability. (b) Will simply increasing the sample size fix Method A's main problem? Explain.

Lời giải chi tiết

(a) Method A has **low variability** (tightly clustered results) but **high bias** (consistently off-center from the true p). Method B has **low bias** (correctly centered on p on average) but **high variability** (widely scattered results). (b) No — increasing the sample size for Method A would only make its results cluster *even more tightly* around the same wrong, too-high value; it would not move that center closer to the true p. Bias is a flaw in the sampling method itself (e.g., a flawed frame or self-selection), and only switching to a better-designed method — not collecting more data with the same flawed method — can fix it.

Probability, Random Variables, and Probability Distributions

Mục tiêu Unit: Ước lượng xác suất bằng mô phỏng và Định luật số lớn (Law of Large Numbers), phân biệt với ngộ nhận của người đánh bạc (gambler's fallacy); nắm vững các quy tắc xác suất cơ bản (quy tắc cộng, biến cố xung khắc, xác suất có điều kiện, quy tắc nhân, tính độc lập, biểu đồ Venn và biểu đồ cây); xây dựng và phân tích phân phối xác suất của biến ngẫu nhiên rời rạc, tính kỳ vọng μ_X và độ lệch chuẩn σ_X; biến đổi và kết hợp các biến ngẫu nhiên độc lập; làm quen với biến ngẫu nhiên liên tục và đường cong mật độ (kể cả phân phối đều); áp dụng phân phối nhị thức (binomial, kể cả xấp xỉ chuẩn) và phân phối hình học (geometric) vào các bài toán thực tế.

4.1 Introducing statistics: estimating probability by simulation

Some sequences of outcomes generated by a genuinely random process look strangely patterned, while some sequences that are actually the product of a hidden pattern can look random at first glance. To avoid being fooled by short-run appearances, statisticians define **probability** formally: the probability of an outcome is the proportion of times it would occur in an indefinitely long series of independent repetitions of a random process – it is never a promise about what happens on any single trial, or even the next several trials.

Because a random process can rarely be repeated infinitely many times in practice, probability is estimated in one of two ways. The **theoretical probability** is calculated directly from a model – for example, counting equally likely outcomes or applying a known probability rule. The **empirical (or simulated) probability** is the observed relative frequency of an outcome after actually running a large number of trials, either physically or with a computer **simulation**.

Simulation is especially valuable when a probability is awkward or effectively impossible to compute directly with the rules developed later in this unit – for instance, an intricate multi-stage process with many possible paths. In such cases the simulated process does the calculating for you, one trial at a time, and the resulting relative frequency serves as a stand-in for the theoretical probability.

People are often poor at telling genuinely random outcomes apart from patterned ones, because of the **gambler's fallacy** – the mistaken belief that a random process has a "memory" and must balance itself out soon (e.g., thinking a coin is "due" for tails after five heads in a row). In reality, independent trials never compensate for the past: after five heads, $P(\text{tails})$ on the next flip is still exactly 0.5.

Ví dụ mẫu · Recognizing genuine randomness

Two students each report a sequence of 20 coin flips. Student 1 reports `HTHTHTHTHTHTHTHTHTHT` (perfect alternation). Student 2 reports a messier sequence that happens to include one run of three tails in a row and one run of three heads in a row, mixed with shorter runs. Which report is more consistent with genuinely flipping a fair coin 20 times?

Solution. Every specific sequence of 20 flips, including the perfectly alternating one, is equally

likely under the fair-coin model – each has probability $(0.5)^{20}$. The key is not the probability of one exact sequence versus another, but which **pattern** the sequence displays. A truly random process has no memory and does not systematically avoid repeats, so runs of three or more identical outcomes show up fairly often across 20 independent flips. A perfectly alternating sequence, with *zero* runs longer than one, is a highly atypical pattern for genuine randomness to produce. People asked to "make up" a random-looking sequence consistently fall for the gambler's fallacy and avoid long runs, producing sequences that alternate suspiciously often – exactly like Student 1's. Student 2's messier sequence, containing ordinary short runs, is far more consistent with genuine independent coin flips.

VN **GIẢI THÍCH TIẾNG VIỆT**

Ngộ nhận của người đánh bạc (gambler's fallacy) là niềm tin sai lầm rằng một quá trình ngẫu nhiên có "trí nhớ" và phải tự cân bằng lại – ví dụ nghĩ rằng sau 5 lần ra mặt ngửa liên tiếp thì lần tiếp theo "phải" ra mặt sấp. Thực tế, các lần thử **độc lập** không hề ảnh hưởng lẫn nhau: xác suất vẫn là 0.5 dù trước đó ra gì. Vì ngộ nhận này, khi con người **tự bịa** một dãy kết quả "trông có vẻ ngẫu nhiên," họ thường tránh các chuỗi lặp dài (run) – trong khi dữ liệu ngẫu nhiên thật sự lại thường xuyên có các chuỗi lặp như vậy.

Law of Large Numbers (LLN): as the number of independent, identically distributed trials n increases, the observed **relative frequency** of an outcome converges to its true **theoretical probability**. This convergence happens only in the long run: the relative frequency after a small number of trials can deviate substantially from the theoretical value, and it need not move closer to that value on every additional trial – only the overall long-run trend is guaranteed.

VN **GIẢI THÍCH TIẾNG VIỆT**

Định luật số lớn (Law of Large Numbers – LLN) phát biểu rằng khi số lần thử n càng lớn, **tần suất tương đối quan sát được** (xác suất thực nghiệm/mô phỏng) sẽ càng gần với **xác suất lý thuyết** thật sự của biến cố. Điểm cần lưu ý: LLN chỉ đúng về **lâu dài** – với n nhỏ, tần suất quan sát có thể lệch khá xa giá trị lý thuyết, và không nhất thiết mỗi lần thử tiếp theo sẽ kéo tần suất lại gần hơn ngay lập tức; chỉ có **xu hướng chung** là hội tụ về xác suất thật khi $n \to \infty$. Đây là lý do vì sao mô phỏng với cỡ mẫu nhỏ có thể cho kết quả gây hiểu lầm.

Designing a simulation

A well-designed simulation used to estimate a probability has four components.

1. **State the assumptions** of the underlying probability model (e.g., each free throw is made independently with probability 0.7).
2. **Assign a random device** to imitate one trial – e.g., let two-digit random numbers 00–69 represent "made" and 70–99 represent "missed," matching the stated probability 0.7, and discard any repeated or invalid digits if needed.
3. **Run many trials**, recording the outcome of each one exactly as the assumptions specify.
4. **Summarize**: compute the relative frequency of the outcome of interest across all trials – this relative frequency is the simulated (empirical) probability estimate.

More trials produce a more reliable estimate, by the Law of Large Numbers.

Ví dụ mẫu · Simulating with a table of random digits

A basketball player makes free throws independently with probability $p = 0.70$ on each attempt. Use a table of random digits to simulate 10 independent attempts, and report the simulated (empirical) probability of a make.

Solution. Step 1 (assumptions): each attempt is independent, with $P(\text{make}) = 0.70$ on every attempt. **Step 2 (assign digits):** let the single digits $0, 1, 2, 3, 4, 5, 6$ represent a make (7 of the 10 equally likely digits, matching $p = 0.70$) and the digits $7, 8, 9$ represent a miss. **Step 3 (run trials):** read the following row of random digits, one digit per attempt:

$$5 \quad 2 \quad 9 \quad 6 \quad 0 \quad 3 \quad 8 \quad 4 \quad 1 \quad 4$$

Attempt	1	2	3	4	5	6	7	8	9	10
Digit	5	2	9	6	0	3	8	4	1	4
Outcome	M	M	X	M	M	M	X	M	M	M

M = make, X = miss.

Step 4 (summarize): only the digits 9 and 8 (attempts 3 and 7) fall in the "miss" range; the remaining 8 attempts are makes. The simulated probability of a make in this particular run is $\hat{p} = \dfrac{8}{10} = 0.80$. This one run of 10 trials landed noticeably above the theoretical value $p = 0.70$ – a reminder that with only 10 trials, relative frequency can deviate substantially from the true probability. Repeating the simulation many more times (or running far more trials per repetition) would, by the Law of Large Numbers, pull the long-run simulated proportion of makes back toward 0.70.

Ví dụ mẫu · Simulating a fair coin

A coin is claimed to be fair, so its theoretical probability of heads on any flip is $p = 0.5$. A simulation of repeated independent flips produces the running results below.

Flips (n)	Cumulative heads	Relative frequency $\hat{p} = \dfrac{\text{heads}}{n}$	$\lvert\hat{p} - 0.5\rvert$
10	6	0.600	0.100
20	11	0.550	0.050
50	24	0.480	0.020
100	53	0.530	0.030
200	97	0.485	0.015
500	252	0.504	0.004
1000	489	0.489	0.011

The relative frequency does not shrink its distance from 0.5 at every single step – it rises from 0.020 back up to 0.030, and later from 0.004 back up to 0.011 – but the overall pattern across the seven rows is a relative frequency staying much closer to the theoretical value $p = 0.5$ as n grows from 10 to 1000, exactly as the Law of Large Numbers predicts. A simulation like this can *support* a claim (here, that the coin is fair), but it can never *prove* it with certainty.

4.2 Introduction to probability: sample space and probability rules

For example, tossing a fair coin twice has sample space $S = \{HH, HT, TH, TT\}$; the event $A =$ "exactly one head" is the subset $\{HT, TH\}$, so if the four outcomes are equally likely, $P(A) = \frac{2}{4} = 0.5$. Not every sample space has equally likely outcomes – a spinner with unequal-sized regions,

for instance – so probabilities must always be read from (or calculated using) the actual model, not assumed equal by default.

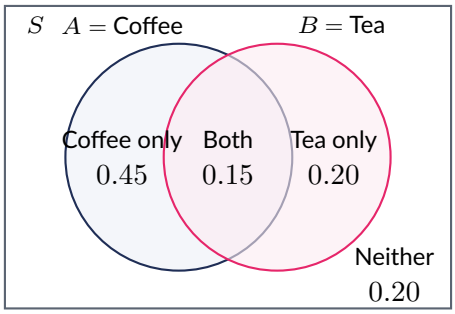
The **sample space** S is the set of all possible outcomes of a random process. An **event** is any subset of the sample space – a collection of outcomes we are interested in. Every probability rule follows from two basic facts: $0 \le P(A) \le 1$ for any event A, and the probabilities of all outcomes in S sum to 1.

Two rules built directly on this foundation are used constantly throughout this unit.

Complement rule: $P(A^c) = P(\text{not } A) = 1 - P(A)$, where A^c ("not A") is every outcome in S that is *not* in A.
Addition rule (general form): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. If A and B are **mutually exclusive** (cannot occur together), then $P(A \cap B) = 0$ and the rule simplifies to $P(A \cup B) = P(A) + P(B)$.

VN **GIẢI THÍCH TIẾNG VIỆT**

Quy tắc phần bù (complement rule): xác suất "A không xảy ra" luôn bằng 1 trừ đi xác suất "A xảy ra". **Quy tắc cộng (addition rule):** để tính $P(A \cup B)$ (A hoặc B xảy ra), cộng $P(A)$ với $P(B)$ rồi **trừ đi phần giao** $P(A \cap B)$ để tránh đếm trùng những kết quả nằm trong cả hai biến cố. Nếu A và B **không thể xảy ra đồng thời** (xung khắc), phần giao bằng 0 nên chỉ cần cộng trực tiếp hai xác suất.



Venn diagram for the coffee/tea example: the four disjoint regions (only A, only B, both, neither) partition the whole sample space S and sum to 1.

Ví dụ mẫu · Addition rule with a Venn diagram

Let A = the event a randomly selected adult drinks coffee, and B = the event the same adult drinks tea. Suppose $P(A) = 0.60$, $P(B) = 0.35$, and $P(A \cap B) = 0.15$ (drinks both). Find $P(A \cup B)$, and use it to find the probability the adult drinks *only* coffee, *only* tea, and *neither*.

Solution. By the addition rule, $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.60 + 0.35 - 0.15 = 0.80$. Coffee only: $P(A) - P(A \cap B) = 0.60 - 0.15 = 0.45$. Tea only: $P(B) - P(A \cap B) = 0.35 - 0.15 = 0.20$. Neither: $1 - P(A \cup B) = 1 - 0.80 = 0.20$. As a check, the four disjoint regions of the Venn diagram sum to 1: $0.45 + 0.15 + 0.20 + 0.20 = 1.00$. So 80% of adults drink coffee, tea, or both, and 20% drink neither beverage.

4.3 Mutually exclusive (disjoint) events

Events A and B are **mutually exclusive** (equivalently, **disjoint**) if they cannot both occur on the same trial: $P(A \cap B) = 0$. On a Venn diagram, mutually exclusive events are drawn as two circles that do not overlap.

A common misconception is to treat "mutually exclusive" as a fancy synonym for "unrelated," in the everyday sense of the word *independent*. In probability, the two ideas are almost opposites: for two events with positive probability, being mutually exclusive is actually the *strongest* possible form of dependence.

Why mutually exclusive events cannot be independent

Suppose A and B are mutually exclusive with $P(A) > 0$ and $P(B) > 0$. Then $P(A \cap B) = 0$. If A and B were also independent, the multiplication rule would require $P(A \cap B) = P(A) \cdot P(B)$ – a product of two *positive* numbers, which cannot equal 0. This is a contradiction, so mutually exclusive events with positive probability are **never** independent. Intuitively: knowing that A occurred tells you *for certain* that B did not occur (since they cannot happen together) – knowledge of A completely changes the probability of B, which is the opposite of independence.

VN **GIẢI THÍCH TIẾNG VIỆT**

Đừng nhầm "xung khắc" (mutually exclusive – không thể cùng xảy ra) với "độc lập" (independent – biến cố này không ảnh hưởng đến xác suất biến cố kia). Trên thực tế, nếu A và B đều có xác suất dương và xung khắc nhau, biết A xảy ra sẽ cho biết **chắc chắn** B không xảy ra – đây là mức độ **phụ thuộc mạnh nhất** có thể có, hoàn toàn trái ngược với tính độc lập.

Ví dụ mẫu · Mutually exclusive events and the addition rule

A dealership's records show that among vehicles sold last month, $P(\text{sedan}) = 0.40$, $P(\text{SUV}) = 0.35$, and $P(\text{truck}) = 0.25$ (every vehicle sold falls into exactly one of these three categories). Find $P(\text{sedan or truck})$.

Solution. A single vehicle cannot be both a sedan and a truck, so the two events are mutually exclusive: $P(\text{sedan} \cap \text{truck}) = 0$. The addition rule simplifies to $P(\text{sedan or truck}) = P(\text{sedan}) + P(\text{truck}) = 0.40 + 0.25 = 0.65$. As a check, the three categories are mutually exclusive *and* exhaustive (every vehicle is exactly one type), so their probabilities sum to 1: $0.40 + 0.35 + 0.25 = 1.00$.

4.4 Conditional probability

4.4.1 Definition and the conditional probability formula

The **conditional probability** of A given that B has occurred is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0.$$

"Given B" means we restrict attention entirely to the outcomes where B has already happened, and ask what fraction of *those* outcomes also satisfy A.

(!) Lỗi thường gặp: Trong công thức $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$, mẫu số **luôn là xác suất của biến cố đứng sau dấu "|"** (biến cố đã biết chắc chắn đã xảy ra), **không phải** $P(A)$. Viết nhầm thành $\dfrac{P(A \cap B)}{P(A)}$ sẽ cho ra $P(B \mid A)$ – một đại lượng hoàn toàn khác với $P(A \mid B)$, trừ khi tình cờ $P(A) = P(B)$.

4.4.2 Two-way tables

Two-way tables display counts (or probabilities) jointly classified by two categorical variables, and are one of the most natural places to compute conditional probabilities directly from row or column totals.

Addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (subtract 0 if mutually exclusive). **Conditional probability:** $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$. **Independence:** A, B are independent iff $P(A \cap B) = P(A) \cdot P(B)$, equivalently $P(A \mid B) = P(A)$.

Ví dụ mẫu · Two-way table

200 students are classified by class year and sport participation:

	Sport	No sport	Total
Freshman	30	20	50
Sophomore	25	35	60
Junior	20	30	50
Senior	15	25	40
Total	90	110	200

$P(\text{Sport}) = \frac{90}{200} = 0.45$. $P(\text{Sport} \mid \text{Freshman}) = \frac{30}{50} = 0.60$. Since $0.60 \neq 0.45$, **Sport and Freshman are not independent.** Addition rule: $P(\text{Junior or Sport}) = P(\text{Junior}) + P(\text{Sport}) - P(\text{Junior and Sport}) = 0.25 + 0.45 - 0.10 = 0.60$.

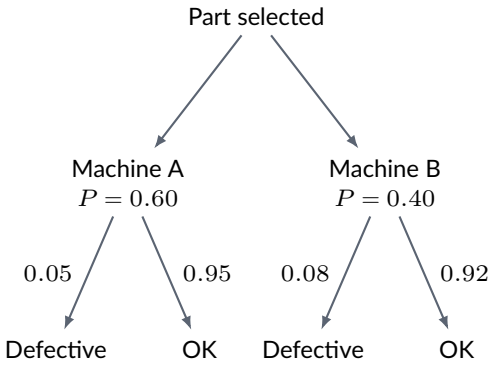
VN **GIẢI THÍCH TIẾNG VIỆT**

Kiểm tra độc lập: so sánh $P(A \mid B)$ với $P(A)$ – nếu bằng nhau thì độc lập. **Lưu ý:** hai biến cố **xung khắc** (mutually exclusive, không thể cùng xảy ra) mà đều có xác suất dương thì **không bao giờ** độc lập, vì biết một biến cố xảy ra sẽ cho biết chắc chắn biến cố kia không xảy ra.

Two-way tables also introduce useful vocabulary for locating a probability inside the table. A **joint probability** is the probability of two categories occurring together, found in a single interior cell of the table (e.g., $P(\text{Freshman and Sport}) = \frac{30}{200} = 0.15$). A **marginal probability** is the probability of just one variable, ignoring the other, found in a row or column total (e.g., $P(\text{Sport}) = \frac{90}{200} = 0.45$, from the "Total" row). A **conditional probability** restricts to a single row or column and divides by *that* row or column's total, as computed above. Every conditional probability in a two-way table can therefore be written as a ratio of a joint probability to a marginal probability, which is exactly the formula $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$.

4.4.3 Tree diagrams and the general multiplication rule

Rearranging the conditional probability formula gives the **general multiplication rule**: $P(A \cap B) = P(A) \cdot P(B \mid A)$. A **tree diagram** organizes a multi-stage random process so this rule can be applied visually: multiply probabilities *along* a branch to get the probability of that path, and add *across* branches that lead to the same final outcome.



Tree diagram for the manufacturing example: multiplying along each path (general multiplication rule) gives $P(A \cap \text{Def}) = 0.60 \times 0.05 = 0.03$ and $P(B \cap \text{Def}) = 0.40 \times 0.08 = 0.032$.

Ví dụ mẫu · Tree diagram and the general multiplication rule

A factory has two machines. Machine A produces 60% of all parts, and 5% of Machine A's parts are defective. Machine B produces the remaining 40% of parts, and 8% of Machine B's parts are defective. A part is selected at random.

(a) Find $P(\text{Machine A and defective})$.

(b) Find $P(\text{defective})$.

(c) Given the part is defective, find $P(\text{Machine A} \mid \text{defective})$.

Solution. Multiply along each branch of the tree:

$$P(A \cap \text{Def}) = P(A) \cdot P(\text{Def} \mid A) = 0.60 \times 0.05 = 0.030$$
$$P(B \cap \text{Def}) = P(B) \cdot P(\text{Def} \mid B) = 0.40 \times 0.08 = 0.032$$

(a) $P(\text{Machine A and defective}) = 0.030$.
(b) Every defective part comes from exactly one machine, so add the two branch probabilities: $P(\text{Def}) = 0.030 + 0.032 = 0.062$.
(c) $P(\text{Machine A} \mid \text{defective}) = \dfrac{P(A \cap \text{Def})}{P(\text{Def})} = \dfrac{0.030}{0.062} \approx 0.4839$. Even though Machine A produces most of the parts and has the lower defect rate, once we know a part is defective there is only about a 48.4% chance it came from Machine A – noticeably lower than the unconditional 60%, because Machine B's higher defect rate pulls a disproportionate share of the defective parts toward it. This process – dividing one branch of a tree by the sum of all branches leading to the observed outcome – is the basic reasoning behind what is formally known as Bayes' theorem.

4.5 Independent events and unions of events

Events A and B are **independent** if knowing one occurred does not change the probability of the other: $P(A \mid B) = P(A)$ (equivalently $P(B \mid A) = P(B)$). This is algebraically equivalent to $P(A \cap B) = P(A) \cdot P(B)$ – the **multiplication rule for independent events**, which is the special case of the general multiplication rule $P(A \cap B) = P(A) \cdot P(B \mid A)$ when $P(B \mid A) = P(B)$.

Independence vs. mutual exclusivity

- **Mutually exclusive:** A and B *cannot* both happen – $P(A \cap B) = 0$. A statement about the *outcomes* of A and B never overlapping.

- **Independent:** whether A happens has *no effect* on the probability that B happens – $P(A \mid B) = P(A)$. A statement about *probabilities not influencing each other*.

As shown above, two events with positive probability that are mutually exclusive can never be independent, and two events that are independent (with positive probability) can never be mutually exclusive. Checking one of these two properties tells you nothing about the other, except that they can never *both* hold at once.

Ví dụ mẫu · Checking independence and applying the multiplication rule

Roll a fair six-sided die once. Let A = "the roll is even" ($P(A) = \frac{1}{2}$) and B = "the roll is at least 5" ($P(B) = \frac{1}{3}$, outcomes $\{5, 6\}$). Are A and B independent?

Solution. $A \cap B$ = "even and at least 5" = $\{6\}$, so $P(A \cap B) = \frac{1}{6}$. Check the multiplication rule: $P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$. Since $P(A \cap B) = P(A) \cdot P(B) = \frac{1}{6}$, A **and** B **are independent** – among the outcomes $\{5, 6\}$, exactly one (6) is even, so the conditional probability of "even" given "at least 5" is still $\frac{1}{2}$, unchanged from the unconditional probability.

VN **GIẢI THÍCH TIẾNG VIỆT**

Hai cách kiểm tra độc lập tương đương nhau: so sánh $P(A \mid B)$ với $P(A)$, hoặc so sánh $P(A \cap B)$ với $P(A) \cdot P(B)$ – chỉ cần một trong hai cách cho kết quả bằng nhau là đủ để kết luận độc lập. Quy tắc nhân tổng quát $P(A \cap B) = P(A) \cdot P(B \mid A)$ luôn đúng cho mọi biến cố; quy tắc nhân đơn giản $P(A \cap B) = P(A) \cdot P(B)$ chỉ đúng khi A và B **độc lập**.

The general multiplication rule is essential precisely when trials are *not* independent – most commonly, when sampling **without replacement** from a small, finite group. Removing an item changes the composition of what remains, so the probability on the second draw depends on the result of the first.

Ví dụ mẫu · General multiplication rule without independence

A bag contains 5 red marbles and 3 blue marbles. Two marbles are drawn *without replacement*. Find $P(\text{both marbles are red})$.

Solution. The first draw: $P(\text{red}_1) = \dfrac{5}{8}$. Because the marble is not returned, only 4 red marbles remain out of 7 total for the second draw: $P(\text{red}_2 \mid \text{red}_1) = \dfrac{4}{7}$. By the general multiplication rule,

$$P(\text{both red}) = P(\text{red}_1) \cdot P(\text{red}_2 \mid \text{red}_1) = \frac{5}{8} \cdot \frac{4}{7} = \frac{20}{56} = \frac{5}{14} \approx 0.357.$$

Notice that $P(\text{red}_2 \mid \text{red}_1) = \frac{4}{7} \neq \frac{5}{8} = P(\text{red}_1)$ – the two draws are **not independent**, so only the *general* multiplication rule (not the simplified independent-events version) applies here.

A very common exam pattern combines independence with the complement rule to find the probability that *at least one* of several independent events occurs – directly computing this by adding individual probabilities is tempting but wrong, since "at least one" includes many overlapping ways for two or more events to happen together, which a simple sum would double-count.

Ví dụ mẫu · "At least one" with independent events

A car has 4 independent smoke detectors, each working correctly with probability 0.98 on a given day. Find the probability that *at least one* of the four detectors fails on a given day.

Solution. It is far easier to work with the complement, "none of the four detectors fails." Since the detectors are independent, the probability that all four work is the product of the individual probabilities:

$$P(\text{none fail}) = P(\text{all 4 work}) = (0.98)^4 \approx 0.92237.$$

By the complement rule, $P(\text{at least one fails}) = 1 - P(\text{none fail}) = 1 - 0.92237 \approx 0.0776$. There is about a 7.8% chance that at least one of the four detectors fails on a given day, even though each individual detector is highly reliable.

(!) Lỗi thường gặp: Với bài toán "**ít nhất một**" (at least one) trong số nhiều biến cố **độc lập**, học sinh thường mắc lỗi **cộng trực tiếp** các xác suất riêng lẻ (ví dụ $P(\text{lỗi 1}) + P(\text{lỗi 2}) + \cdots$), dẫn đến **đếm trùng** các trường hợp nhiều biến cố cùng xảy ra. Cách làm đúng và nhanh nhất luôn là dùng **quy tắc phần bù**: $P(\text{ít nhất một xảy ra}) = 1 - P(\text{không có biến cố nào xảy ra})$, và với các biến cố độc lập, $P(\text{không có biến cố nào xảy ra})$ là **tích** của các xác suất "không xảy ra" riêng lẻ.

4.6 Introduction to random variables and probability distributions

A **discrete** random variable takes values that can be listed (finitely many, or countably infinite, like $0, 1, 2, 3, \ldots$), each with a positive probability – this is the type of random variable studied for most of this unit. A **continuous** random variable instead takes any value in an interval of real numbers (such as height or time), so probability is described by area under a density curve rather than by a table; continuous random variables are developed later in this unit, after the discrete case.

By convention, an uppercase letter such as X denotes the random variable itself (the process of generating a numerical outcome), while a lowercase letter such as x denotes one specific possible value it could take. For example, if X = the number of heads in 3 tosses of a fair coin, then X can take the values $x = 0, 1, 2, 3$, and $P(X = 2)$ asks for the probability that this particular random process produces the specific value 2.

A **random variable** assigns a numerical value to each outcome of a random process. A **discrete random variable** X takes a countable list of possible values $x_1, x_2, \ldots$, each with its own probability. Its **probability distribution** is a table (or formula) listing every possible value of X together with $P(X = x_i)$. A table is a **valid probability distribution** exactly when (1) every probability satisfies $0 \le P(x_i) \le 1$, and (2) the probabilities sum to exactly 1: $\sum_i P(x_i) = 1$.

VN **GIẢI THÍCH TIẾNG VIỆT**

Biến ngẫu nhiên rời rạc (discrete random variable) nhận một danh sách các giá trị đếm được, mỗi giá trị gắn với một xác suất. Một bảng phân phối xác suất được gọi là **hợp lệ** khi thỏa hai điều: mọi xác suất đều nằm trong khoảng $[0, 1]$, và **tổng tất cả các xác suất phải bằng đúng** 1. Nếu tổng khác 1, hoặc có một xác suất âm hay lớn hơn 1, bảng đó **không phải** là một phân phối xác suất hợp lệ.

Ví dụ mẫu · Recognizing a valid probability distribution

A discrete random variable X has the proposed distribution below. Is it a valid probability distribution?

x	0	1	2	3
$P(X = x)$	0.20	0.45	0.30	0.10

Solution. Every listed probability lies between 0 and 1. Sum: $0.20+0.45+0.30+0.10 = 1.05 \neq 1$. Because the probabilities sum to 1.05, not 1, this is **not** a valid probability distribution – at least one value must be corrected before this table can describe a real random variable.

4.7 Mean and standard deviation of random variables

For discrete random variable X: $\mu_X = E(X) = \sum x_i P(x_i)$, and $\sigma_X^2 = \sum (x_i - \mu_X)^2 P(x_i)$, with $\sigma_X = \sqrt{\sigma_X^2}$. The mean μ_X is the long-run average value of X over many repetitions; the standard deviation σ_X measures the typical distance of X from that average.

Ví dụ mẫu · Building a distribution and computing μ_X, σ_X

A carnival game costs $\$4$ to play. A player spins a wheel: with probability 0.10 they win a $\$20$ prize, with probability 0.30 they win a $\$5$ prize, and with probability 0.60 they win nothing. Let X = the player's *net* profit (prize won minus the $\$4$ cost). Build the probability distribution of X, then find μ_X and σ_X.

Solution. Net profit: win $\$20$ prize $\rightarrow X = 20 - 4 = 16$; win $\$5$ prize $\rightarrow X = 5 - 4 = 1$; win nothing $\rightarrow X = 0 - 4 = -4$.

x	$P(x)$	$xP(x)$	$(x-\mu_X)^2$	$(x-\mu_X)^2 P(x)$
16	0.10	1.6	272.25	27.225
1	0.30	0.3	2.25	0.675
-4	0.60	-2.4	12.25	7.350
Sum	1.00	-0.5		35.250

$\mu_X = \sum xP(x) = 1.6 + 0.3 - 2.4 = -0.5$. Using $\mu_X = -0.5$ to build the last column: $\sigma_X^2 = \sum (x-\mu_X)^2 P(x) = 35.25$, so $\sigma_X = \sqrt{35.25} \approx 5.94$. On average, a player loses about $\$0.50$ per play (the game favors the house), and any individual play's net profit typically differs from that average by about $\$5.94$, due to the wide range of possible prizes.

Notice that $\mu_X = -0.5$ is not itself a possible outcome of a single play – the game only ever pays 16, 1, or -4 dollars. Like most averages, the expected value is a long-run summary describing what happens *on average across many repetitions*, not a value that must occur on any individual trial.

4.8 Transforming and combining random variables

Transforming: $E(aX+b) = aE(X) + b$, $\text{Var}(aX+b) = a^2\text{Var}(X)$ (so $\sigma_{aX+b} = |a|\sigma_X$).
Combining independent X, Y: $E(X \pm Y) = E(X) \pm E(Y)$, and $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ (variances *always add* for independent variables, even when subtracting).

(!) Lỗi thường gặp: Khi **cộng hoặc trừ** hai biến ngẫu nhiên độc lập, phương sai luôn **cộng** (không bao giờ trừ), vì cả hai nguồn biến thiên đều làm kết quả kém chắc chắn hơn: $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$.

VN **GIẢI THÍCH TIẾNG VIỆT**
Khi **nhân** biến ngẫu nhiên với hằng số a rồi **cộng** hằng số b: kỳ vọng biến đổi theo cùng công thức $aE(X)+b$, nhưng phương sai chỉ bị nhân với a^2 (hằng số cộng b không ảnh hưởng đến độ

trải). Khi **kết hợp hai biến độc lập**, kỳ vọng cộng/trừ bình thường như số thực, nhưng phương sai (không phải độ lệch chuẩn) **luôn cộng lại**, dù ta đang cộng hay trừ hai biến.

Ví dụ mẫu · Linear transformation of a random variable

Average daily high temperatures in a city during a certain month are modeled by a random variable X (in °C) with $\mu_X = 18$ and $\sigma_X = 4$. Convert to degrees Fahrenheit using $Y = \frac{9}{5}X + 32$. Find μ_Y and σ_Y, and interpret each.

Solution. Applying the transformation rules with $a = \frac{9}{5}$ and $b = 32$:

$$\mu_Y = \frac{9}{5}\mu_X + 32 = \frac{9}{5}(18) + 32 = 32.4 + 32 = 64.4\ ^\circ\text{F}, \qquad \sigma_Y = \left|\frac{9}{5}\right|\sigma_X = \frac{9}{5}(4) = 7.2\ ^\circ\text{F}.$$

The additive constant 32 shifts every value (and therefore the mean) upward but does not stretch or compress the distribution, so it contributes nothing to σ_Y. The multiplicative constant $\frac{9}{5}$ rescales the entire distribution, so it multiplies *both* the mean and the standard deviation. On average, the daily high is 64.4°F, typically varying from day to day by about 7.2°F.

Ví dụ mẫu · Combining two independent random variables

At a coffee shop, the time to take an order has mean $\mu_X = 3.2$ minutes and SD $\sigma_X = 0.8$ minutes. The time to prepare the drink is independent of the order time, with mean $\mu_Y = 2.5$ minutes and SD $\sigma_Y = 0.6$ minutes. Let $T = X + Y$ be the total time, and $D = X - Y$ the difference between the two stage times.

Solution. Means: $\mu_T = \mu_X + \mu_Y = 3.2 + 2.5 = 5.7$ minutes; $\mu_D = \mu_X - \mu_Y = 3.2 - 2.5 = 0.7$ minutes. Because X and Y are independent, variances add *in both cases*: $\text{Var}(T) = \text{Var}(D) = \sigma_X^2 + \sigma_Y^2 = 0.8^2 + 0.6^2 = 0.64 + 0.36 = 1.00$, so $\sigma_T = \sigma_D = \sqrt{1.00} = 1.0$ minute. On average, a customer's total wait is 5.7 minutes, with the order stage typically taking 0.7 minutes longer than the preparation stage; both the total time and the difference in stage times vary by about 1.0 minute from customer to customer.

4.9 Continuous random variables and density curves

So far, every random variable in this unit has been discrete, taking a countable list of separated values. Many quantities of real interest – height, time, weight, a proportion – instead take *any* value in an interval of real numbers, and are described not by a probability table but by a smooth **density curve**.

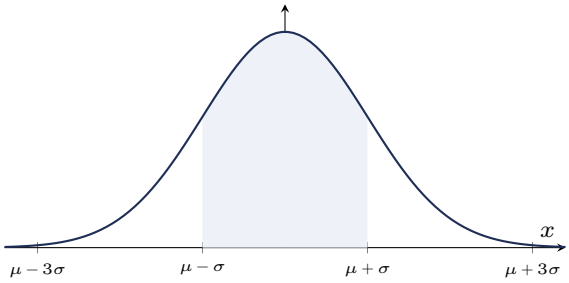
A **continuous random variable** X is described by a **density curve**: a curve that lies on or above the horizontal axis, with **total area under the curve equal to** 1. The probability that X falls in an interval equals the **area under the density curve above that interval**: $P(a < X < b) =$ the area under the curve between $x = a$ and $x = b$. Because a single point has zero width, the area above any single value is 0, so for a continuous random variable $P(X = a) = 0$ for every a – consequently $P(X < a) = P(X \le a)$, since including or excluding a single endpoint never changes the probability.

VN **GIẢI THÍCH TIẾNG VIỆT**

Với biến ngẫu nhiên **liên tục** (continuous random variable), xác suất được biểu diễn bằng **diện tích dưới đường cong mật độ** (density curve), chứ không phải bằng một bảng liệt kê từng giá trị như biến rời rạc. Tổng diện tích dưới toàn bộ đường cong luôn bằng 1. Vì một điểm đơn

lẻ không có "bề rộng" nên diện tích tại đúng một điểm bằng 0 – do đó $P(X = a) = 0$ với biến liên tục, và việc lấy dấu "$<$" hay "$\leq$" **không làm thay đổi** xác suất.

The most important continuous density curve in this course is the **Normal curve**: symmetric, single-peaked, and bell-shaped, fully determined by its mean μ and standard deviation σ. The **68–95–99.7 rule** gives quick area (probability) estimates for any Normal curve: about 68% of the total area lies within 1σ of μ, about 95% within 2σ, and about 99.7% within 3σ.



The Normal density curve: the shaded area between $\mu - \sigma$ and $\mu + \sigma$ represents $P(\mu - \sigma < X < \mu + \sigma) \approx 0.68$, matching the 68–95–99.7 rule. Total area under any density curve equals 1.

Ví dụ mẫu · Area under the Normal density curve

Heights of adult men in a population are approximately Normal with $\mu = 70$ inches and $\sigma = 2.5$ inches. Using the 68–95–99.7 rule, find (a) $P(65 < X < 75)$ and (b) $P(X > 72.5)$.

Solution. (a) $65 = 70 - 2(2.5)$ and $75 = 70 + 2(2.5)$, so this interval spans exactly $\mu \pm 2\sigma$: $P(65 < X < 75) \approx 0.95$.
(b) $72.5 = 70 + 1(2.5) = \mu + \sigma$. About 68% of the area lies within 1σ of μ (between 67.5 and 72.5), leaving $1 - 0.68 = 0.32$ split evenly between the two symmetric tails, so each tail holds about $\frac{0.32}{2} = 0.16$: $P(X > 72.5) \approx 0.16$.

4.9.1 The uniform distribution

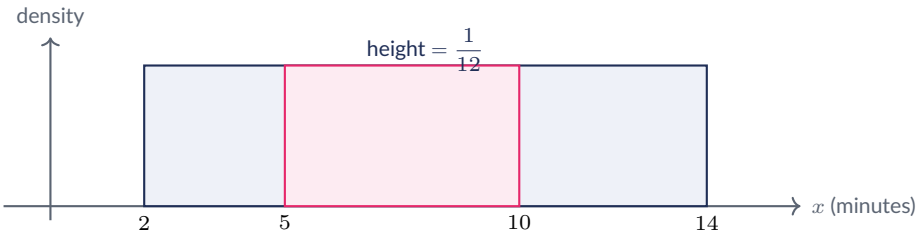
A **uniform distribution** is the simplest continuous model: its density curve is a flat, constant height over an interval $[a, b]$ (and 0 outside it), so every sub-interval of a given length is equally likely, regardless of where it falls within $[a, b]$.

A continuous random variable X is **uniformly distributed** on $[a, b]$ if its density curve has the constant height $\dfrac{1}{b-a}$ for $a \leq x \leq b$ (and height 0 elsewhere), so the total area is $(b-a) \times \dfrac{1}{b-a} = 1$. For any $c < d$ within $[a, b]$, the probability of landing between them is simply the area of a rectangle:

$$P(c < X < d) = (d - c) \times \frac{1}{b-a}.$$

VN **GIẢI THÍCH TIẾNG VIỆT**

Phân phối đều (uniform distribution) trên đoạn $[a, b]$ có đường cong mật độ là một **đoạn thẳng nằm ngang** với chiều cao không đổi $\dfrac{1}{b-a}$ – nghĩa là mọi khoảng con có cùng độ dài đều có xác suất bằng nhau, bất kể vị trí của nó trong $[a, b]$. Vì đường cong là một hình chữ nhật, xác suất $P(c < X < d)$ chỉ đơn giản là **diện tích hình chữ nhật** có chiều dài $(d - c)$ và chiều cao $\dfrac{1}{b-a}$.



Uniform density curve for browsing time on $[2, 14]$ minutes (height not drawn to scale): total area $= 12 \times \frac{1}{12} = 1$. The shaded region, from 5 to 10 minutes, has area $P(5 < T < 10) = 5 \times \frac{1}{12} = \frac{5}{12} \approx 0.417$.

Ví dụ mẫu · The uniform distribution

A city bus is scheduled to arrive every 20 minutes. In practice, its arrival time W (minutes after the scheduled time) is uniformly distributed between 0 and 20 minutes. Find $P(5 < W < 12)$ and $P(W < 5)$.

Solution. The density height is $\dfrac{1}{20 - 0} = 0.05$.

$$P(5 < W < 12) = (12{-}5){\times}0.05 = 7{\times}0.05 = 0.35, \qquad P(W < 5) = (5{-}0){\times}0.05 = 5{\times}0.05 = 0.25.$$

There is a 35% chance the bus arrives between 5 and 12 minutes late, and a 25% chance it arrives within the first 5 minutes of the scheduled time.

4.10 The binomial distribution

4.10.1 Introducing the binomial distribution

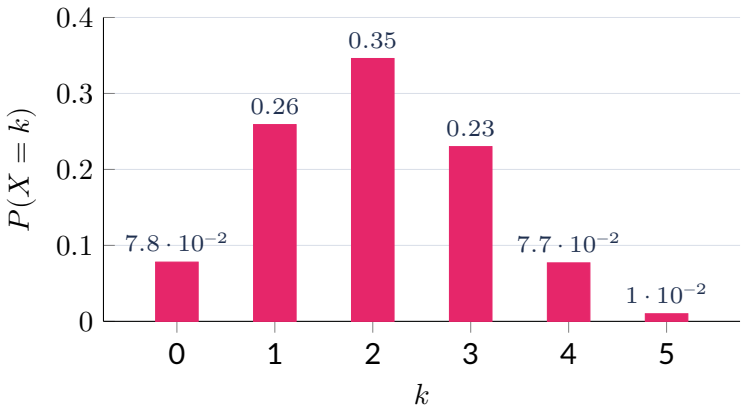
The binomial setting arises naturally whenever a fixed number of identical, independent yes/no trials are repeated – inspecting parts off an assembly line, surveying independent shoppers, or recording makes and misses across a fixed number of shot attempts. The formula for $P(X = k)$ has two parts, each answering a different question: the binomial coefficient $\binom{n}{k}$ counts *how many different arrangements* of k successes and $n - k$ failures are possible among the n trials, while $p^k(1-p)^{n-k}$ is the probability of *any one* such specific arrangement (successes and failures are independent, so probabilities along a single arrangement multiply). Since every arrangement with exactly k successes has this same probability, multiplying the two together gives the total probability of exactly k successes in any order.

A **binomial** setting requires: **B**inary outcomes (each trial is success/failure), **I**ndependent trials, a fixed **N**umber of trials n, and the **S**ame probability p of success on every trial (BINS). Under these conditions, if X = the number of successes in n trials, then

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \dots, n.$$

VN **GIẢI THÍCH TIẾNG VIỆT**

Bốn điều kiện BINS cần kiểm tra trước khi dùng công thức nhị thức: **B**inary (kết quả nhị phân – thành công/thất bại), **I**ndependent (các lần thử độc lập với nhau), **N** cố định (số lần thử n không đổi), **S**ame p (xác suất thành công p giống nhau ở mọi lần thử). Thiếu một trong bốn điều kiện, công thức nhị thức **không áp dụng chính xác được**.



Binomial probability distribution, $n = 5$, $p = 0.4$: $\mu = np = 2.0$, $\sigma = \sqrt{np(1-p)} \approx 1.10$.

Ví dụ mẫu · Binomial probability and constructing a distribution table

A basketball player makes free throws independently with $p = 0.4$ each. Out of $n = 5$ attempts, find $P(X = 3)$ made shots, then construct the full distribution table of X.

$$P(X=3) = \binom{5}{3}(0.4)^3(0.6)^2 = 10(0.064)(0.36) = 0.2304.$$

k	0	1	2	3	4
$P(X=k)$	0.078	0.259	0.346	0.230	0.077

(and $P(X=5) = 0.4^5 \approx 0.010$)

Mean: $\mu_X = np = 5(0.4) = 2.0$ makes; $\sigma_X = \sqrt{5(0.4)(0.6)} = \sqrt{1.2} \approx 1.10$.

4.10.2 Parameters, cumulative probabilities, and the 10% condition

For binomial X with parameters n and p: $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$. A **cumulative binomial probability** adds several individual terms: $P(X \le k)$ ("at most k"), $P(X \ge k)$ ("at least k"), and $P(X > k)$ ("more than k," which excludes k itself, unlike "at least"). The complement rule is often the fastest route: $P(X \ge k) = 1 - P(X \le k-1)$.

(!) Lỗi thường gặp: "**Ít nhất** k" (at least k, tức $P(X \ge k)$) và "**nhiều hơn** k" (more than k, tức $P(X > k)$) **khác nhau**: "ít nhất k" **bao gồm** giá trị k, còn "nhiều hơn k" thì **không**. Với $k = 3$ chẳng hạn: $P(X \ge 3)$ cộng cả số hạng $P(X = 3)$, còn $P(X > 3)$ chỉ cộng từ $P(X = 4)$ trở lên. Đọc nhầm một chữ trong đề bài sẽ dẫn tới sai đáp số hoàn toàn.

Ví dụ mẫu · Cumulative binomial probability with the complement rule

On a multiple-choice quiz with 4 choices per question, a student guesses randomly and independently on all $n = 6$ questions, so $p = 0.25$ per question. Find the probability the student gets *at least* 3 questions correct.

Solution. It is faster to use the complement: $P(X \ge 3) = 1 - P(X \le 2) = 1 - [P(0) + P(1) +$

$P(2)]$.

$$P(0) = \binom{6}{0}(0.25)^0(0.75)^6 = 0.75^6 \approx 0.17798$$
$$P(1) = \binom{6}{1}(0.25)^1(0.75)^5 = 6(0.25)(0.23730) \approx 0.35596$$
$$P(2) = \binom{6}{2}(0.25)^2(0.75)^4 = 15(0.0625)(0.31641) \approx 0.29663$$

$P(X \le 2) \approx 0.17798 + 0.35596 + 0.29663 = 0.83057$. So $P(X \ge 3) = 1 - 0.83057 \approx 0.1694$. There is about a 16.9% chance the student gets at least 3 of the 6 questions right by guessing alone.

When sampling **without replacement** from a finite population, the probability of success technically changes slightly after each draw – strictly speaking, this violates the "independent trials" and "same p" conditions of BINS. In practice, when the sample is small relative to the population, this change is negligible and the binomial model still gives an excellent approximation.

10% condition: when sampling without replacement, trials may be treated as approximately independent (and p as approximately constant) provided the sample size satisfies $n \le 0.10N$, where N is the population size. If $n > 0.10N$, removing individuals changes the remaining proportion of successes too much for the binomial model to be trustworthy.

Ví dụ mẫu · Checking the 10% condition

A quality control manager samples $n = 15$ items without replacement to inspect. Should she treat the number of defective items as approximately binomial if the shipment contains (a) $N = 120$ items, or (b) $N = 2000$ items?

Solution. (a) $0.10 \times 120 = 12$, and $n = 15 > 12$, so the 10% condition **fails**: sampling 15 out of only 120 items changes the remaining composition of the shipment too much from draw to draw, so the binomial model should **not** be used. (b) $0.10 \times 2000 = 200$, and $n = 15 \le 200$, so the condition **holds**: sampling 15 out of 2000 barely changes the remaining proportion of defectives, so treating the count of defectives as approximately binomial is reasonable.

(!) Lỗi thường gặp: Đừng quên kiểm tra điều kiện 10% khi lấy mẫu **không hoàn lại** (without replacement) từ một tổng thể **hữu hạn**. Nếu cỡ mẫu n vượt quá 10% kích thước tổng thể N, xác suất thành công p sẽ thay đổi **đáng kể** sau mỗi lần chọn, vi phạm giả định "p không đổi" của phân phối nhị thức – khi đó công thức nhị thức chỉ là một xấp xỉ kém và không nên áp dụng.

4.10.3 Normal approximation to the binomial distribution

When n is large, computing an exact binomial cumulative probability by summing many individual terms becomes impractical by hand. Because a binomial distribution with a large n is roughly symmetric and bell-shaped (provided p is not too close to 0 or 1), it can be approximated by a Normal curve with the same mean and standard deviation.

Normal approximation to the binomial: if X is binomial with parameters n and p, and both $np \geq 10$ **and** $n(1-p) \geq 10$, then X is approximately Normal with $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$. Probabilities for X can then be estimated using $z = \dfrac{x - np}{\sqrt{np(1-p)}}$ together with the standard Normal curve.

VN **GIẢI THÍCH TIẾNG VIỆT**

Khi n lớn, việc cộng nhiều số hạng nhị thức bằng tay rất mất công, nên ta có thể **xấp xỉ phân phối nhị thức bằng phân phối chuẩn** có cùng $\mu = np$ và $\sigma = \sqrt{np(1-p)}$ – **với điều kiện** cả $np \geq 10$ **và** $n(1-p) \geq 10$ (đảm bảo phân phối không quá lệch về một phía). Thiếu một trong hai điều kiện, xấp xỉ chuẩn có thể sai lệch đáng kể so với giá trị nhị thức thật, đặc biệt khi p quá gần 0 hoặc 1.

Ví dụ mẫu · Normal approximation to the binomial

A city survey finds that 40% of residents support a proposed transit tax. In an independent random sample of $n = 150$ residents, let X = the number who support it. Estimate $P(X \leq 54)$ using the Normal approximation, and compare with the exact binomial value.

Solution. Check conditions: $np = 150(0.4) = 60 \geq 10$ and $n(1-p) = 150(0.6) = 90 \geq 10$, so the Normal approximation is appropriate. The approximating Normal curve has $\mu_X = 60$ and $\sigma_X = \sqrt{150(0.4)(0.6)} = \sqrt{36} = 6$.

$$z = \frac{54 - 60}{6} = -1.00 \qquad \Rightarrow \qquad P(X \leq 54) \approx P(Z \leq -1.00) = 0.1587.$$

The exact binomial probability (summing $P(X = 0)$ through $P(X = 54)$) is 0.1799 – reasonably close to the Normal approximation, with the small remaining gap coming from the binomial distribution's discreteness. (Using the refinement $z = \frac{54.5-60}{6} \approx -0.92$, called a **continuity correction**, gives $P(Z \leq -0.92) \approx 0.1797$, even closer to the exact value.) The Normal approximation is a genuine approximation, not an exact method – it is most accurate when both np and $n(1-p)$ are comfortably above 10.

4.11 The geometric distribution

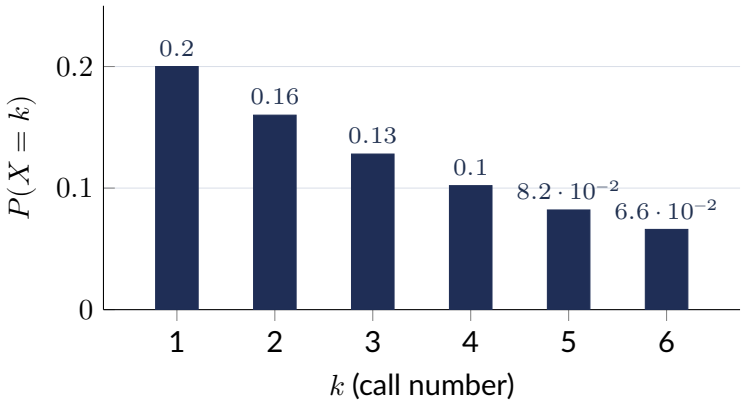
A **geometric** setting shares the **B**inary, **I**ndependent, **S**ame-p conditions with the binomial setting (BIS), but counts the number of trials *up to and including the first success* – there is no fixed number of trials. If X = the trial number of the first success,

$$P(X = k) = (1-p)^{k-1}p, \qquad k = 1, 2, 3, \ldots, \qquad \mu_X = \frac{1}{p}.$$

VN **GIẢI THÍCH TIẾNG VIỆT**

Khác với phân phối nhị thức, phân phối hình học (geometric) **không có** số lần thử cố định – ta đếm số lần thử **cho đến khi thành công lần đầu tiên** thì dừng lại. Giá trị nhỏ nhất có thể của X là 1 (thành công ngay lần thử đầu tiên), **không bao giờ** là 0. Kỳ vọng $\mu_X = \dfrac{1}{p}$: xác suất thành công p càng nhỏ, trung bình càng phải chờ lâu mới thành công.

Aside: the geometric distribution also has a variance formula, $\sigma_X^2 = \dfrac{1-p}{p^2}$ (so $\sigma_X = \sqrt{(1-p)/p^2}$). This is rarely needed in practice – only the mean formula $\mu_X = \dfrac{1}{p}$ is treated as an essential exam-ready formula in this course.



Geometric probability distribution for the sales-call scenario below, $p = 0.20$: probabilities decay geometrically, $P(X = k) = (0.8)^{k-1}(0.2)$.

Ví dụ mẫu · Geometric probability

A fair die is rolled repeatedly. Find the probability that the *first* 6 appears on the 4th roll.

Here $p = \frac{1}{6}$ (geometric). $P(X = 4) = \left(1 - \frac{1}{6}\right)^3 \left(\frac{1}{6}\right) = \left(\frac{5}{6}\right)^3 \left(\frac{1}{6}\right) = \frac{125}{216} \cdot \frac{1}{6} = \frac{125}{1296} \approx 0.0965.$

Ví dụ mẫu · Cumulative geometric probability

A recruiter contacts job candidates one at a time; each independently accepts an offer with probability $p = 0.20$. Find (a) the probability that *more than* 5 candidates must be contacted before the first acceptance, and (b) the probability that *at most* 5 candidates must be contacted.

Solution. "More than 5 trials needed" means the first 5 trials *all* fail: $P(X > 5) = (1-p)^5 = (0.8)^5 = 0.32768$. By the complement rule, $P(X \le 5) = 1 - P(X > 5) = 1 - 0.32768 = 0.67232$. So there is about a 32.8% chance the recruiter needs more than 5 contacts to get an acceptance, and about a 67.2% chance the first acceptance happens within the first 5 contacts.

(!) Lỗi thường gặp: Với biến ngẫu nhiên hình học, $P(X > k) = (1-p)^k$ (**không phải** $(1-p)^{k-1}$), vì "nhiều hơn k lần thử" nghĩa là **cả k lần thử đầu tiên đều thất bại**. Ngoài ra, X luôn bắt đầu từ 1 – không tồn tại $P(X = 0)$ trong phân phối hình học, vì cần ít nhất một lần thử để "thành công lần đầu tiên" có thể xảy ra.

Binomial vs. geometric: how to tell them apart

Both settings share the same **B**inary, **I**ndependent, **S**ame-p conditions, but ask a different question about the trials.

- **Binomial:** the number of trials n is **fixed in advance**; the random variable X counts the number of successes *within* those n trials. X can equal $0, 1, 2, \dots, n$.
- **Geometric:** the number of trials is **not fixed** – trials continue until the first success occurs; the random variable X counts *how many trials that takes*. X can equal $1, 2, 3, \dots$, with no

fixed upper limit.

A quick test while reading a problem: if it specifies "out of n trials" or gives a fixed sample size, suspect binomial; if it asks "how many trials until ..." or "the probability the first ...occurs on trial ...," suspect geometric.

Formula reference for Unit 4

Rule / quantity	Formula
Complement	$P(A^c) = 1 - P(A)$
Addition rule	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Conditional probability	$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
General multiplication rule	$P(A \cap B) = P(A) \cdot P(B \mid A)$
Independence check	$P(A \cap B) = P(A)P(B)$, equivalently $P(A \mid B) = P(A)$
Discrete RV mean	$\mu_X = \sum x_i P(x_i)$
Discrete RV variance	$\sigma_X^2 = \sum (x_i - \mu_X)^2 P(x_i)$
Transforming	$E(aX + b) = aE(X) + b$, $\text{Var}(aX + b) = a^2\text{Var}(X)$
Combining independent X, Y	$E(X \pm Y) = E(X) \pm E(Y)$, $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$
Continuous RV	$P(a < X < b) =$ area under density curve from a to b; total area $= 1$
Uniform on $[a, b]$	$P(c < X < d) = (d - c) \times \dfrac{1}{b - a}$
Binomial probability	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$; $\mu_X = np$, $\sigma_X = \sqrt{np(1-p)}$
Normal approx. to binomial	valid if $np \geq 10$ and $n(1-p) \geq 10$; use $\mu = np$, $\sigma = \sqrt{np(1-p)}$
Geometric probability	$P(X = k) = (1-p)^{k-1} p$; $\mu_X = \dfrac{1}{p}$; $P(X > k) = (1-p)^k$

4.12 Practice problems

Bài 1

A survey shows 30% of customers prefer brand X. Among $n = 8$ randomly and independently chosen customers, find $P(\text{exactly } 2 \text{ prefer brand } X)$.

(A) 0.0576

(B) 0.1765

(C) 0.2965

(D) 0.3000

Lời giải chi tiết

Binomial with $n = 8$, $p = 0.3$: $P(X = 2) = \binom{8}{2}(0.3)^2(0.7)^6 = 28(0.09)(0.117649) \approx 0.2965$.

(C).

Bài 2

A random variable X has $\mu_X = 50$, $\sigma_X = 5$. Let $Y = 3X - 10$. Find μ_Y and σ_Y.

(A) $\mu_Y = 140,\ \sigma_Y = 15$
(C) $\mu_Y = 140,\ \sigma_Y = 5$
(B) $\mu_Y = 150,\ \sigma_Y = 5$
(D) $\mu_Y = 140,\ \sigma_Y = 45$

Lời giải chi tiết

$\mu_Y = 3(50) - 10 = 140$. Adding a constant does not affect spread, and multiplying scales the SD by $|3|$: $\sigma_Y = 3(5) = 15$. **(A)**.

Bài 3

A spinner lands on "win" with probability $p = 0.25$ on each independent spin. Find the probability that the first win occurs on exactly the 3rd spin.

(A) 0.140625
(C) 0.25
(B) 0.046875
(D) 0.5625

Lời giải chi tiết

Geometric: $P(X = 3) = (1 - 0.25)^2(0.25) = (0.75)^2(0.25) = 0.5625(0.25) = 0.140625$. **(A)**.

Bài 4

In a large population, 44% of people have type O blood. What is the probability that a randomly selected person does *not* have type O blood?

(A) 0.44
(C) 1.44
(B) 0.56
(D) 0.22

Lời giải chi tiết

By the complement rule, $P(\text{not O}) = 1 - P(\text{O}) = 1 - 0.44 = 0.56$. **(B)**.

Bài 5

Two fair six-sided dice are rolled, and S is the sum of the two faces shown.

(a) How many outcomes are in the sample space?

(b) Find $P(S = 7)$.

(c) Find $P(S \geq 10)$.

Lời giải chi tiết

(a) Each die has 6 faces, and the rolls are independent, so the sample space has $6 \times 6 = 36$ equally likely outcomes.
(b) Sum $= 7$: $(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)$ – 6 outcomes, so $P(S = 7) = \frac{6}{36} = \frac{1}{6} \approx 0.167$.
(c) Sum ≥ 10 means sum $= 10, 11$, or 12: $(4,6), (5,5), (6,4)$ for 10; $(5,6), (6,5)$ for 11; $(6,6)$ for 12 – 6 outcomes total, so $P(S \geq 10) = \frac{6}{36} = \frac{1}{6} \approx 0.167$.

Bài 6

A single card is drawn from a standard, well-shuffled 52-card deck. Let A = "the card is a King" and B = "the card is a Queen." Find $P(A \text{ or } B)$.

(A) 0.1538 **(C)** 0.3077

(B) 0.0769 **(D)** 0.0059

Lời giải chi tiết

A card cannot be both a King and a Queen, so A and B are mutually exclusive: $P(A \cap B) = 0$. Addition rule: $P(A \cup B) = P(A) + P(B) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13} \approx 0.1538$. **(A)**.

Bài 7

At a school, 62% of students take Spanish, 28% take French, and 15% take both languages. Find the probability that a randomly selected student takes Spanish or French, and the probability that a student takes neither.

Lời giải chi tiết

Addition rule: $P(\text{Spanish or French}) = P(S) + P(F) - P(S \cap F) = 0.62 + 0.28 - 0.15 = 0.75$. By the complement rule, $P(\text{neither}) = 1 - 0.75 = 0.25$. So 75% of students take at least one of the two languages, and 25% take neither.

Bài 8

Let A = the event a randomly selected registered voter is a Democrat, and B = the event the same voter is a Republican (assume no voter is registered as both). Suppose $P(A) = 0.35$ and $P(B) = 0.30$. Explain whether A and B are mutually exclusive, and whether they are independent.

Lời giải chi tiết

Since no voter can be registered as both a Democrat and a Republican, A and B **are mutually exclusive**: $P(A \cap B) = 0$. To check independence, compare $P(A \cap B) = 0$ to $P(A) \cdot P(B) = 0.35 \times 0.30 = 0.105$. Since $0 \neq 0.105$, A and B are **not independent**. This illustrates the general fact that mutually exclusive events with positive probabilities can never be independent: knowing a voter is a Democrat tells you with certainty that they are not a Republican, so the two probabilities strongly affect each other.

Bài 9

A survey of 250 adults classifies them by exercise habits and daily vitamin use:

	Vitamin	No vitamin	Total
Exercise	70	50	120
No exercise	40	90	130
Total	110	140	250

Find $P(\text{Vitamin} \mid \text{Exercise})$.

(A) 0.583
(B) 0.280
(C) 0.440
(D) 0.636

Lời giải chi tiết

Restrict attention to the "Exercise" row (120 people total), of whom 70 take a vitamin: $P(\text{Vitamin} \mid \text{Exercise}) = \dfrac{70}{120} \approx 0.583$. **(A)**. (D) is the reversed conditional probability $P(\text{Exercise} \mid \text{Vitamin}) = \frac{70}{110} \approx 0.636$.

Bài 10

A shipment of electronic components comes 60% from Supplier X and 40% from Supplier Y. Historically, 3% of Supplier X's components are defective, while 7% of Supplier Y's components are defective. A component is chosen at random from the shipment.

(a) Find $P(\text{Supplier Y and defective})$.

(b) Find $P(\text{defective})$.

(c) Given the component is defective, find $P(\text{Supplier X} \mid \text{defective})$.

Lời giải chi tiết

Using the general multiplication rule along each tree branch:

$$P(X \cap \text{Def}) = 0.60 \times 0.03 = 0.018, \qquad P(Y \cap \text{Def}) = 0.40 \times 0.07 = 0.028.$$

(a) $P(\text{Supplier Y and defective}) = 0.028$.
(b) $P(\text{Def}) = 0.018 + 0.028 = 0.046$.
(c) $P(\text{Supplier X} \mid \text{Def}) = \dfrac{0.018}{0.046} \approx 0.3913$. Even though Supplier X provides most of the components (60%), only about 39.1% of the *defective* components come from Supplier X, since Supplier Y's higher defect rate contributes a disproportionate share of the defective pile.

Bài 11

Using the same table as the previous problem (exercise vs. vitamin use), what can you conclude about the independence of "Exercise" and "Vitamin use"?

(A) They are independent, since both probabilities are positive
(B) They are not independent, since $P(\text{Vitamin} \mid \text{Exercise}) \neq P(\text{Vitamin})$
(C) They are mutually exclusive
(D) Independence cannot be determined from a two-way table

Lời giải chi tiết

$P(\text{Vitamin}) = \dfrac{110}{250} = 0.44$. From problem 9, $P(\text{Vitamin} \mid \text{Exercise}) = \dfrac{70}{120} \approx 0.583$. Since $0.583 \neq 0.44$, exercise status changes the probability of taking a vitamin, so the two variables are **not independent**. **(B)**.

Bài 12

Suppose $P(A) = 0.6$ and $P(B \mid A) = 0.25$. Find $P(A \cap B)$. If it is also known that $P(B) = 0.3$, are A and B independent?

Lời giải chi tiết

General multiplication rule: $P(A \cap B) = P(A) \cdot P(B \mid A) = 0.6 \times 0.25 = 0.15$. To check independence, compare to $P(A) \cdot P(B) = 0.6 \times 0.3 = 0.18$. Since $P(A \cap B) = 0.15 \neq 0.18 = P(A)P(B)$, A and B are **not independent** (equivalently, $P(B \mid A) = 0.25 \neq 0.30 = P(B)$, confirming the same conclusion).

Bài 13

Which statement about two events A and B, each with positive probability, is TRUE?

(A) If A and B are mutually exclusive, they are also independent.

(B) If A and B are independent, they can still be mutually exclusive.

(C) If A and B are mutually exclusive, they cannot be independent.

(D) Mutually exclusive and independent describe the same relationship.

Lời giải chi tiết

For events with positive probability, mutual exclusivity forces $P(A \cap B) = 0$, while independence would require $P(A \cap B) = P(A)P(B) > 0$ – these cannot both hold. **(C)**.

Bài 14

A discrete random variable X has distribution $P(X = 0) = 0.15$, $P(X = 1) = 0.35$, $P(X = 2) = k$, $P(X = 3) = 0.10$. Find the value of k that makes this a valid probability distribution.

(A) 0.40

(B) 0.60

(C) 0.25

(D) 0.15

Lời giải chi tiết

All probabilities must sum to 1: $0.15 + 0.35 + k + 0.10 = 1 \Rightarrow 0.60 + k = 1 \Rightarrow k = 0.40$. **(A)**.

Bài 15

A discrete random variable X has the distribution below.

x	1	2	3	4
$P(X = x)$	0.1	0.4	0.3	0.2

Find μ_X and σ_X.

Lời giải chi tiết

$\mu_X = \sum xP(x) = 1(0.1) + 2(0.4) + 3(0.3) + 4(0.2) = 0.1 + 0.8 + 0.9 + 0.8 = 2.6$.

$\sigma_X^2 = \sum(x-\mu_X)^2 P(x)$ with $\mu_X = 2.6$:

$$\begin{aligned}\sigma_X^2 &= (1-2.6)^2(0.1) + (2-2.6)^2(0.4) + (3-2.6)^2(0.3) + (4-2.6)^2(0.2)\\ &= 0.256 + 0.144 + 0.048 + 0.392 = 0.84.\end{aligned}$$

So $\sigma_X = \sqrt{0.84} \approx 0.917$.

Bài 16

Exam scores in a class have $\mu_X = 80$ and $\sigma_X = 12$ (out of 100). The teacher rescales every score using $Y = 0.8X + 20$. Find μ_Y and σ_Y.

(A) $\mu_Y = 84,\ \sigma_Y = 12$

(B) $\mu_Y = 100,\ \sigma_Y = 9.6$

(C) $\mu_Y = 84,\ \sigma_Y = 9.6$

(D) $\mu_Y = 84,\ \sigma_Y = 29.6$

Lời giải chi tiết

$\mu_Y = 0.8(80) + 20 = 64 + 20 = 84$. The additive constant 20 does not affect spread; the multiplicative constant scales SD: $\sigma_Y = 0.8(12) = 9.6$. **(C)**.

Bài 17

At a grocery store, the time to scan a customer's items has mean 4.5 minutes and SD 1.2 minutes. The time to process payment is independent, with mean 1.8 minutes and SD 0.5 minutes. Find the mean and SD of the total checkout time.

Lời giải chi tiết

Mean of a sum: $\mu_{\text{total}} = 4.5 + 1.8 = 6.3$ minutes. Because the two stages are independent, variances add: $\sigma^2_{\text{total}} = 1.2^2 + 0.5^2 = 1.44 + 0.25 = 1.69$, so $\sigma_{\text{total}} = \sqrt{1.69} = 1.3$ minutes. Total checkout time averages 6.3 minutes, typically varying by about 1.3 minutes from customer to customer.

Bài 18

Two independent baristas prepare drinks. Barista A's time has mean 3.5 minutes and SD 0.9 minutes; Barista B's time has mean 3.0 minutes and SD 1.2 minutes. Find the mean and SD of the difference $A - B$ in their times.

(A) $\mu = 0.5,\ \sigma = 1.5$

(B) $\mu = 0.5,\ \sigma = 0.3$

(C) $\mu = 0.5,\ \sigma = 2.25$

(D) $\mu = 6.5,\ \sigma = 1.5$

Lời giải chi tiết

$\mu_{A-B} = 3.5 - 3.0 = 0.5$ minutes. Variances add even for a difference, since both baristas are independent sources of variability: $\sigma^2_{A-B} = 0.9^2 + 1.2^2 = 0.81 + 1.44 = 2.25$, so $\sigma_{A-B} = \sqrt{2.25} = 1.5$ minutes. **(A)**. (B) incorrectly subtracts the SDs directly ($1.2 - 0.9 = 0.3$), which is never valid.

Bài 19

Which of the following does *not* satisfy the conditions for a binomial random variable?

(A) Flipping a fair coin 20 times, counting the heads

(B) Drawing 5 cards without replacement, counting red cards

(C) Surveying 50 independent shoppers, each with buy probability 0.3

(D) 10 independent free throws, each made with probability 0.7

Lời giải chi tiết

Drawing cards *without replacement* means the probability of a red card changes after each draw, and the trials are not independent – this violates both the "Independent" and "Same p" conditions of BINS. **(B)**. The other three scenarios all satisfy binary outcomes, independence, a fixed number of trials, and a constant success probability.

Bài 20

An airline reports that 90% of its flights depart on time. For a random sample of $n = 6$ independent flights, find the probability that exactly 5 depart on time.

Lời giải chi tiết

Binomial with $n = 6$, $p = 0.9$, $k = 5$:

$$P(X = 5) = \binom{6}{5}(0.9)^5(0.1)^1 = 6(0.59049)(0.1) = 0.354294 \approx 0.3543.$$

Bài 21

A factory inspects $n = 200$ independently produced units, each defective with probability $p = 0.15$. Find the mean and standard deviation of the number of defective units.

(A) $\mu = 30,\ \sigma \approx 5.05$

(B) $\mu = 30,\ \sigma \approx 25.5$

(C) $\mu = 170,\ \sigma \approx 5.05$

(D) $\mu = 30,\ \sigma \approx 4.5$

Lời giải chi tiết

$\mu_X = np = 200(0.15) = 30$. $\sigma_X = \sqrt{np(1-p)} = \sqrt{200(0.15)(0.85)} = \sqrt{25.5} \approx 5.05$. **(A)**.

Bài 22

A vaccine successfully protects against a disease in 85% of recipients, independently. Among $n = 5$ vaccinated people exposed to the disease, find the probability that *at least* 4 are protected.

(A) 0.8352

(B) 0.3915

(C) 0.4437

(D) 0.1648

Lời giải chi tiết

Binomial, $n = 5$, $p = 0.85$: $P(X \geq 4) = P(X = 4) + P(X = 5)$.

$$P(X = 4) = \binom{5}{4}(0.85)^4(0.15)^1 = 5(0.52200625)(0.15) \approx 0.39150$$
$$P(X = 5) = (0.85)^5 \approx 0.44371$$

$P(X \geq 4) \approx 0.39150 + 0.44371 = 0.83521$. **(A).**

Bài 23

A fair coin is flipped $n = 10$ independent times. Find $P(\text{at most } 2 \text{ heads})$.

(A) 0.0547
(B) 0.1719
(C) 0.9453
(D) 0.0322

Lời giải chi tiết

Binomial, $n = 10$, $p = 0.5$: $P(X \leq 2) = P(0) + P(1) + P(2)$.

$$P(0) = \binom{10}{0}(0.5)^{10} = \tfrac{1}{1024} \approx 0.00098$$
$$P(1) = \binom{10}{1}(0.5)^{10} = \tfrac{10}{1024} \approx 0.00977$$
$$P(2) = \binom{10}{2}(0.5)^{10} = \tfrac{45}{1024} \approx 0.04395$$

Sum: $P(X \leq 2) \approx 0.00098 + 0.00977 + 0.04395 = 0.0547$. **(A).**

Bài 24

A committee samples $n = 25$ students without replacement to survey opinions on a new policy. Should the number of students favoring the policy be treated as approximately binomial if the school has (a) $N = 200$ students, or (b) $N = 3000$ students? Justify using the 10% condition.

Lời giải chi tiết

Compare n to $0.10N$. (a) $0.10(200) = 20$; since $n = 25 > 20$, the condition **fails** – do not treat as approximately binomial. (b) $0.10(3000) = 300$; since $n = 25 \leq 300$, the condition **holds** – the binomial model is a reasonable approximation.

Bài 25

A basketball player's free-throw attempts are independent, each made with probability 0.75. Find the probability that her first *miss* occurs on the 4th attempt.

(A) 0.1055
(B) 0.3164
(C) 0.0039
(D) 0.4219

Lời giải chi tiết

Let "success" mean a miss, with $p = 1 - 0.75 = 0.25$. The first three attempts must be makes (probability 0.75 each), then a miss: $P(X = 4) = (0.75)^3(0.25) = 0.421875(0.25) = 0.10546875 \approx 0.1055$. **(A)**.

Bài 26

Using the same free-throw shooter as the previous problem ($p = 0.25$ probability of a miss on each attempt), find the average number of attempts needed until her first miss.

Lời giải chi tiết

For a geometric random variable, $\mu_X = \dfrac{1}{p} = \dfrac{1}{0.25} = 4$ attempts.

Bài 27

A recruiter contacts candidates independently; each accepts an offer with probability $p = 0.30$. Find (a) the probability that more than 4 candidates must be contacted before the first acceptance, and (b) the probability that at most 3 candidates are needed.

Lời giải chi tiết

(a) "More than 4" means the first 4 contacts all fail: $P(X > 4) = (1 - 0.30)^4 = (0.7)^4 = 0.2401$.
(b) $P(X \leq 3) = 1 - (0.7)^3 = 1 - 0.343 = 0.657$.

Bài 28

A factory machine produces defective parts independently with probability $p = 0.04$. In a random sample of $n = 150$ parts, find the mean and standard deviation of the number of defective parts, and interpret both values in context.

Lời giải chi tiết

This is binomial with $n = 150$, $p = 0.04$ (BINS conditions are reasonable if parts are produced independently). $\mu_X = np = 150(0.04) = 6$ parts. $\sigma_X = \sqrt{np(1-p)} = \sqrt{150(0.04)(0.96)} = \sqrt{5.76} = 2.4$ parts. **Interpretation:** in batches of 150 parts, the factory expects an average of 6 defective parts, with the actual count in a given batch typically varying from this average by about 2.4 parts.

Bài 29

A school trip's total cost is the sum of two independent components: bus rental, with mean $\$450$ and SD $\$30$, and meals for the group, with mean $\$720$ and SD $\$50$.

(a) Find the mean and SD of the total trip cost.

(b) A $\$50$ discount coupon is applied, reducing the total cost by a fixed $\$50$. Find the new mean and SD.

Lời giải chi tiết

(a) Mean of the sum: $\mu_T = 450 + 720 = 1170$. Since the two costs are independent, variances add: $\sigma_T^2 = 30^2 + 50^2 = 900 + 2500 = 3400$, so $\sigma_T = \sqrt{3400} \approx 58.31$. Total cost averages $\$1170$, typically varying by about $\$58.31$.
(b) Subtracting the fixed $\$50$ coupon shifts the mean but not the spread: $\mu_{T-50} = 1170 - 50 = 1120$, while $\sigma_{T-50} = \sigma_T \approx 58.31$ (unchanged, since subtracting a constant does not affect variability).

Bài 30

A screening test for a disease has sensitivity $P(\text{positive} \mid \text{disease}) = 0.95$ and specificity $P(\text{negative} \mid \text{no disease}) = 0.90$. In the tested population, 2% actually have the disease.

(a) Using a tree-diagram approach, find $P(\text{positive test})$.

(b) Find $P(\text{disease} \mid \text{positive test})$.

(c) If 5 people from this population are tested independently, find the probability that at least 1 tests positive.

Lời giải chi tiết

Tree branches, multiplying along each path: $P(\text{disease}) = 0.02$, $P(\text{pos} \mid \text{disease}) = 0.95$, so $P(\text{disease} \cap \text{pos}) = 0.02(0.95) = 0.019$. Also $P(\text{no disease}) = 0.98$ and $P(\text{pos} \mid \text{no disease}) = 1 - 0.90 = 0.10$ (complement of specificity), so $P(\text{no disease} \cap \text{pos}) = 0.98(0.10) = 0.098$.
(a) Add the two branches leading to a positive test: $P(\text{positive}) = 0.019 + 0.098 = 0.117$.
(b) $P(\text{disease} \mid \text{positive}) = \dfrac{0.019}{0.117} \approx 0.1624$ – even with a fairly accurate test, only about 16.2% of positive results are true disease cases, because the disease is rare (only 2% prevalence) and the pool of healthy false positives is large.
(c) Treat the 5 independent tests as binomial with $p = P(\text{positive}) = 0.117$. $P(\text{at least } 1 \text{ positive}) = 1 - P(\text{none positive}) = 1 - (1 - 0.117)^5 = 1 - (0.883)^5 \approx 1 - 0.5368 = 0.4632$.

Bài 31

The number of siblings X for students in a large school survey has the distribution below.

x	0	1	2	3	4
$P(X = x)$	0.25	0.40	0.20	0.10	0.05

Find μ_X and σ_X.

Lời giải chi tiết

$\mu_X = \sum xP(x) = 0(0.25) + 1(0.40) + 2(0.20) + 3(0.10) + 4(0.05) = 0 + 0.40 + 0.40 + 0.30 + 0.20 = 1.30$ siblings.
$\sigma_X^2 = \sum (x - \mu_X)^2 P(x)$ with $\mu_X = 1.30$:

$$\sigma_X^2 = (0 - 1.3)^2(0.25) + (1 - 1.3)^2(0.40) + (2 - 1.3)^2(0.20) + (3 - 1.3)^2(0.10) + (4 - 1.3)^2(0.05)$$
$$= 0.4225 + 0.036 + 0.098 + 0.289 + 0.3645 = 1.21.$$

So $\sigma_X = \sqrt{1.21} = 1.10$ siblings.

Bài 32

The number of days per week X that an employee at a company works overtime has the distribution below.

x	0	1	2	3	4
$P(X=x)$	0.30	0.25	0.20	0.15	0.10

Find μ_X and σ_X, and interpret μ_X in context.

Lời giải chi tiết

$\mu_X = 0(0.30) + 1(0.25) + 2(0.20) + 3(0.15) + 4(0.10) = 0 + 0.25 + 0.40 + 0.45 + 0.40 = 1.50$ days.

$\sigma_X^2 = (0-1.5)^2(0.30)+(1-1.5)^2(0.25)+(2-1.5)^2(0.20)+(3-1.5)^2(0.15)+(4-1.5)^2(0.10) = 0.675+0.0625+0.05-$

$\sigma_X = \sqrt{1.75} \approx 1.32$ days. On average, an employee works overtime about 1.5 days per week, typically varying from this average by about 1.32 days.

Bài 33

A large orchard reports that 20% of its apple trees show early signs of a fungal infection. In a random sample of $n = 100$ independently inspected trees, let X = the number showing signs of infection.

(a) Verify that X may be approximated by a Normal distribution.

(b) Use the Normal approximation to estimate $P(X \le 15)$.

Lời giải chi tiết

(a) $np = 100(0.20) = 20 \ge 10$ and $n(1-p) = 100(0.80) = 80 \ge 10$, so the Normal approximation is appropriate.
(b) $\mu_X = np = 20$, $\sigma_X = \sqrt{np(1-p)} = \sqrt{100(0.20)(0.80)} = \sqrt{16} = 4$.

$$z = \frac{15-20}{4} = -1.25 \qquad \Rightarrow \qquad P(X \le 15) \approx P(Z \le -1.25) \approx 0.1056.$$

(For reference, the exact binomial value is about 0.1285; the Normal approximation lands in the right neighborhood but is noticeably less precise here than in a case with larger np and $n(1-p)$.)

Bài 34

The time T (in minutes) that a customer spends browsing before checkout at a small shop is uniformly distributed between 2 and 14 minutes.

(a) Find the height of the density curve.

(b) Find $P(5 < T < 10)$.

(A) 0.417
(B) 0.357
(C) 0.500
(D) 0.083

Lời giải chi tiết

(a) Height $= \dfrac{1}{14-2} = \dfrac{1}{12} \approx 0.0833$.
(b) $P(5 < T < 10) = (10-5) \times \dfrac{1}{12} = \dfrac{5}{12} \approx 0.417$. **(A)**.

Bài 35

A network has 3 independent backup servers, each failing on a given day with probability 0.05. Find the probability that *at least one* of the three servers fails on a given day.

(A) 0.1426
(B) 0.15
(C) 0.0001
(D) 0.8574

Lời giải chi tiết

Use the complement: $P(\text{at least one fails}) = 1 - P(\text{none fail})$. By independence, $P(\text{none fail}) = (0.95)^3 = 0.857375$, so $P(\text{at least one fails}) = 1 - 0.857375 = 0.142625 \approx 0.1426$. **(A)**. Choice (B) is the incorrect "add the probabilities" shortcut ($3 \times 0.05 = 0.15$), which overcounts and is never the correct method.

Bài 36

A discrete random variable X has $P(X=-2) = 0.30$, $P(X=0) = 0.45$, $P(X=5) = 0.25$.

(a) Verify this is a valid probability distribution.

(b) Find μ_X and σ_X.

Lời giải chi tiết

(a) All three probabilities lie in $[0,1]$, and $0.30 + 0.45 + 0.25 = 1.00$, so this **is** a valid probability distribution.
(b) $\mu_X = (-2)(0.30) + 0(0.45) + 5(0.25) = -0.6 + 0 + 1.25 = 0.65$.

$\sigma_X^2 = (-2-0.65)^2(0.30) + (0-0.65)^2(0.45) + (5-0.65)^2(0.25) = 2.10675 + 0.190125 + 4.730625 = 7.0275$.

$\sigma_X = \sqrt{7.0275} \approx 2.65$.

Sampling Distributions

Mục tiêu Unit: Phân biệt tham số (parameter) và thống kê mẫu (statistic); hiểu sai số chọn mẫu (sampling variability) và tự tay xây dựng phân phối lấy mẫu (sampling distribution) từ một tổng thể nhỏ; phân biệt tính không thiên lệch (unbiasedness – tâm của phân phối) với độ biến thiên (variability – độ trải, phụ thuộc cỡ mẫu và cách lấy mẫu, không phụ thuộc tính thiên lệch); tính $\mu_{\hat{p}}$, $\sigma_{\hat{p}}$ và kiểm tra điều kiện Large Counts; tính $\mu_{\bar{x}}$, $\sigma_{\bar{x}}$ và áp dụng Định lý giới hạn trung tâm (CLT) kể cả khi tổng thể lệch; hiểu vì sao $\sigma_{\bar{x}}$ và $\sigma_{\hat{p}}$ giảm theo $1/\sqrt{n}$ chứ không phải $1/n$; làm quen với phân phối lấy mẫu của hiệu hai thống kê độc lập, nền tảng cho các Unit suy diễn (Unit 6–8).

Suppose two students each take a random sample of 50 classmates and ask what proportion own a gaming console. Student A's sample gives $\hat{p} = 0.58$; Student B's sample gives $\hat{p} = 0.64$. Both followed the exact same correct random sampling procedure, yet their answers differ. Neither student made an error — every random sample of a given size, drawn by the same method, is simply one outcome out of an enormous number of samples that *could have* been drawn. Understanding the entire collection of outcomes a statistic can take, and how that collection behaves, is the subject of this unit.

Three features of that collection matter, and they organize everything that follows: its **center** (is the statistic centered at the true parameter, or systematically off?), its **spread** (how much do values of the statistic typically vary from sample to sample, and what controls that variation?), and its **shape** (can a Normal curve be used to compute probabilities about the statistic, and under what conditions?). Every later unit in this course — every confidence interval, every significance test — is built directly on top of the ideas developed here.

5.1 Statistics, parameters, and sampling variability

Khái niệm cốt lõi

A **parameter** describes the population (e.g., μ, p) and is usually unknown. A **statistic** describes a sample (e.g., $\bar{x}$, $\hat{p}$) and varies from sample to sample — this is **sampling variability**. The **sampling distribution** of a statistic is the distribution of its values over all possible samples of a given size. An estimator is **unbiased** if the mean of its sampling distribution equals the true parameter.

VN **GIẢI THÍCH TIẾNG VIỆT**

Thống kê mẫu (statistic, ví dụ $\bar{x}$, $\hat{p}$) thay đổi theo từng mẫu — đó là **biến thiên do lấy mẫu**. "Không thiên lệch" (unbiased) nghĩa là **trung bình** của phân phối lấy mẫu bằng đúng tham số thật; tăng cỡ mẫu n chỉ làm **giảm biến thiên** (độ trải hẹp lại), **không** làm giảm thiên lệch nếu phương pháp lấy mẫu vốn đã lệch.

Every parameter has a natural statistic used to estimate it, and the notation itself is a helpful memory aid: population quantities are almost always Greek letters (μ, p, σ), while the corresponding sample statistics either reuse the Latin letter (s for σ) or add a "hat" or bar to the Greek/Latin symbol ($\bar{x}$ for μ, $\hat{p}$ for p) to signal "this is an estimate, computed from data."

Population quantity	Parameter	Sample quantity	Statistic
Mean	μ	Sample mean	$\bar{x}$
Proportion	p	Sample proportion	$\hat{p}$
Standard deviation	σ	Sample standard deviation	s

A parameter is a single fixed number — it does not change unless the population itself changes. A statistic, by contrast, is a random variable: before the sample is drawn, we do not know what value it will take, and afterward, a different SRS of the same size would almost certainly give a different value. The sampling distribution describes exactly how that randomness behaves — its center, its spread, and its shape — and every confidence interval and significance test built in later units rests directly on the sampling distribution of some statistic.

A useful habit when reading any problem in this unit is to ask, in order: what is the population, what parameter am I being asked about, what is the sample, and what statistic estimates that parameter? Only once these four pieces are correctly identified does it make sense to ask about the statistic's sampling distribution at all.

5.1.1 Building a sampling distribution by hand

The definition of a sampling distribution ("the distribution of a statistic over *all possible* samples of a given size") can feel abstract. For a small enough population, we can construct the entire sampling distribution explicitly and see exactly what the definition means. Real populations are far too large to enumerate every possible sample by hand – that is precisely why formulas such as $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ exist, so we never have to. But working through one small, fully enumerable case first makes those formulas far more than something to memorize.

Ví dụ mẫu · Constructing a sampling distribution from scratch

Consider the tiny population $\{2, 4, 6, 8\}$ ($N = 4$). Its population mean is

$$\mu = \frac{2+4+6+8}{4} = \frac{20}{4} = 5,$$

and its population standard deviation is

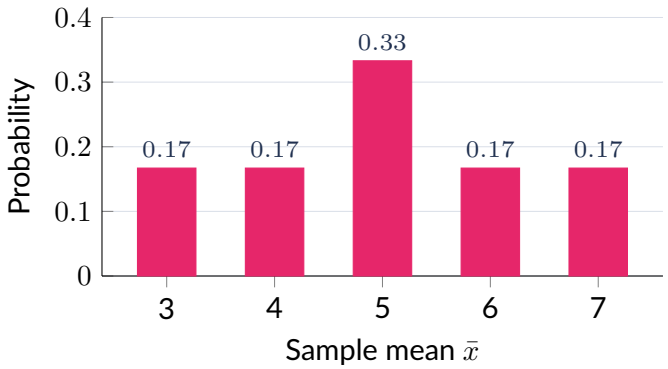
$$\sigma = \sqrt{\frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4}} = \sqrt{\frac{9+1+1+9}{4}} = \sqrt{5} \approx 2.236.$$

Now take an SRS of size $n = 2$, *without replacement*. There are $\binom{4}{2} = 6$ equally likely samples, each with probability $\frac{1}{6}$. Listing every one and its sample mean $\bar{x}$:

Sample	$\{2,4\}$	$\{2,6\}$	$\{2,8\}$	$\{4,6\}$	$\{4,8\}$	$\{6,8\}$
$\bar{x}$	3	4	5	5	6	7

Collecting these six equally-likely outcomes into a distribution table gives the sampling distribution of $\bar{x}$ for $n = 2$:

$\bar{x}$	3	4	5	6	7
Probability	$1/6$	$1/6$	$2/6$	$1/6$	$1/6$



The sampling distribution of $\bar{x}$ for all $\binom{4}{2} = 6$ possible SRSs of size $n = 2$ from the population $\{2, 4, 6, 8\}$.

The mean of this sampling distribution is

$$\mu_{\bar{x}} = 3\left(\tfrac{1}{6}\right) + 4\left(\tfrac{1}{6}\right) + 5\left(\tfrac{2}{6}\right) + 6\left(\tfrac{1}{6}\right) + 7\left(\tfrac{1}{6}\right) = \frac{3+4+10+6+7}{6} = \frac{30}{6} = 5,$$

which is exactly the population mean $\mu = 5$ — direct, hand-computed confirmation that $\bar{x}$ is an **unbiased** estimator of μ: it holds not just approximately, but exactly, once every possible sample is accounted for. The spread of the sampling distribution ($\bar{x}$ ranges only from 3 to 7, clustered near 5) is also visibly tighter than the spread of the population itself (which ranges from 2 to 8): the sampling distribution's standard deviation here is $\sqrt{\frac{(3-5)^2+(4-5)^2+2(0)+(6-5)^2+(7-5)^2}{6}} = \sqrt{\frac{10}{6}} \approx 1.291$, noticeably smaller than $\sigma \approx 2.236$. Note this is *not* exactly $\sigma/\sqrt{n} = 2.236/\sqrt{2} \approx 1.581$: the formula $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is exact only when sampling with replacement, or when the population is large relative to n (the **10% condition**) — here, with $n = 2$ sampled without replacement from a population of only $N = 4$, that condition is badly violated, so a small finite-population adjustment is needed. In virtually every AP Statistics scenario the 10% condition holds comfortably, so this adjustment can safely be ignored and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is used directly.

What happens if the sample size grows, still within the same tiny population? Take $n = 3$ instead of $n = 2$ from $\{2, 4, 6, 8\}$. There are only $\binom{4}{3} = 4$ possible SRSs:

Sample	$\{2, 4, 6\}$	$\{2, 4, 8\}$	$\{2, 6, 8\}$	$\{4, 6, 8\}$
$\bar{x}$	4	$14/3 \approx 4.67$	$16/3 \approx 5.33$	6

The mean of these four values is $\frac{4 + 4.67 + 5.33 + 6}{4} = \frac{20}{4} = 5$ — matching $\mu = 5$ exactly, once again. But the *spread* has shrunk: the standard deviation of this $n = 3$ sampling distribution is $\sqrt{\frac{(4-5)^2+(4.67-5)^2+(5.33-5)^2+(6-5)^2}{4}} = \sqrt{\frac{1+0.111+0.111+1}{4}} = \sqrt{0.556} \approx 0.745$, noticeably smaller than the $n = 2$ spread of ≈ 1.291 computed above. Both sampling distributions are centered at exactly the same place ($\mu = 5$); only the spread changed as n grew from 2 to 3. This small, fully hand-verified comparison is a preview of the general rule developed later in this unit: increasing the sample size never changes the center of a sampling distribution, but it always shrinks the spread. It is the same $1/\sqrt{n}$ relationship that governs every sampling distribution discussed in this chapter, made visible here through direct enumeration in a population small enough to check by hand.

GIẢI THÍCH TIẾNG VIỆT VN

Với một tổng thể nhỏ, ta có thể liệt kê **toàn bộ** các mẫu SRS cỡ n có thể có, tính $\bar{x}$ cho từng mẫu, rồi gom lại thành bảng phân phối lấy mẫu — đây chính là cách "nhìn thấy" định nghĩa trừu tượng của sampling distribution bằng con số cụ thể. Khi lấy trung bình có trọng số của tất cả các $\bar{x}$ này (theo xác suất từng mẫu), ta luôn được đúng μ của tổng thể — đó là bằng chứng trực tiếp cho tính **không thiên lệch**. Đồng thời, độ trải của bảng phân phối này (từ 3 đến 7) hẹp hơn

nhiều so với độ trải của chính tổng thể (từ 2 đến 8) — minh hoạ trực quan cho việc lấy trung bình mẫu làm giảm biến thiên. Khi cỡ mẫu tăng từ 2 lên 3 (vẫn trong tổng thể chỉ có 4 phần tử), độ trải càng hẹp hơn nữa trong khi tâm vẫn không đổi – đúng như quy luật $1/\sqrt{n}$ sẽ được phát biểu chính thức ở phần sau.

5.2 From individual values to sample statistics: a z-score refresher

Before developing the sampling distributions of $\hat{p}$ and $\bar{x}$ in detail, it is worth pausing on a distinction that causes frequent errors: a z-score for a single **individual** observation is a different calculation from a z-score for a **statistic**.

Ví dụ mẫu · Individual z-score – no sample involved

Scores on a national placement exam are approximately Normal with $\mu = 500$ and $\sigma = 100$. What proportion of individual test-takers score above 650?

$$z = \frac{650 - 500}{100} = 1.50. \qquad P(X > 650) = P(Z > 1.50) = 1 - 0.9332 = 0.0668.$$

About 6.68% of individual test-takers score above 650. This calculation uses the population standard deviation σ directly — no sample and no sampling distribution is involved at all; it is a single random individual being compared to the whole population.

Now change the question: instead of one test-taker, take an SRS of $n = 25$ test-takers and ask about their *average* score $\bar{x}$. Averages are far less variable than individual scores — an unusually high or low single score gets diluted by the other 24 scores it is averaged with. Answering this new kind of question requires a different (smaller) standard deviation than σ: the standard deviation of the sampling distribution of $\bar{x}$, developed fully below.

A quick way to decide which calculation a problem requires: if the question asks about a *single randomly chosen individual* ("What is the probability a randomly selected test-taker scores above 650?"), use σ directly, exactly as in Unit 1. If the question asks about a *sample average, sample proportion, or any other statistic* computed from multiple observations ("What is the probability the average of 25 test-takers exceeds 650?"), the standard deviation of *that statistic's* sampling distribution must be used instead — $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ for a mean, or $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ for a proportion. Both are developed in full in the sections that follow. Watch, too, for the wording that signals which case applies: phrases like "a randomly selected," "one," or "an individual" point to σ; phrases like "the sample mean," "the average of," or "the proportion of the sample" point to a sampling-distribution standard deviation instead.

VN **GIẢI THÍCH TIẾNG VIỆT**

Ở Unit 1, ta tính z-score cho **một cá thể** bằng cách dùng trực tiếp σ của **tổng thể**. Nhưng khi câu hỏi chuyển sang **trung bình của cả một mẫu** $\bar{x}$ (thay vì một cá thể), $\bar{x}$ sẽ **ít biến thiên hơn nhiều** so với từng cá thể riêng lẻ, vì các giá trị cao/thấp bất thường bị "trung hoà" khi lấy trung bình. Vì vậy công thức z-score cho $\bar{x}$ (hay $\hat{p}$) phải dùng một độ lệch chuẩn khác – không phải σ, mà là $\sigma_{\bar{x}}$ hoặc $\sigma_{\hat{p}}$ (luôn nhỏ hơn σ của một cá thể). Nhầm lẫn hai đại lượng này – dùng σ của cá thể để tính xác suất cho một thống kê – là lỗi cực kỳ phổ biến khi làm bài thi.

5.3 Bias and variability of a statistic

The hand-built example above showed unbiasedness directly: the mean of the sampling distribution matched the population mean exactly. That property — where the sampling distribution's center lands — is only one of two properties that matter for judging how good a statistic is as an estimator. The other, entirely separate, property is how spread out that sampling distribution is.

Bias concerns the **center** of a sampling distribution: an estimator is **unbiased** if $\mu_{\text{statistic}} =$ parameter, and **biased** if its sampling distribution is systematically too high or too low. **Variability** concerns the **spread** of a sampling distribution, and is controlled by the sample size n and the sampling design — *not* by whether the estimator is biased. These are two completely separate properties: a statistic can be unbiased yet highly variable, or biased yet very precise (low variability). A well-designed study aims for both properties at once – an estimator that is both unbiased *and* has low variability – but achieving one does not guarantee the other.

VN **GIẢI THÍCH TIẾNG VIỆT**

Bias liên quan đến **tâm** của phân phối lấy mẫu (giá trị trung bình của thống kê có đúng bằng tham số thật hay không), còn **độ biến thiên** (variability) liên quan đến **độ trải** của phân phối đó, và độ trải này phụ thuộc vào **cỡ mẫu** n và **cách thiết kế lấy mẫu** – KHÔNG phụ thuộc vào việc thống kê có bị thiên lệch hay không. Một ước lượng có thể không thiên lệch nhưng vẫn rất biến thiên (mẫu nhỏ), hoặc bị thiên lệch nhưng lại rất ổn định (sai một cách nhất quán) – đây là hai tính chất độc lập với nhau.

5.3.1 A biased estimator: the sample maximum

Not every natural-seeming statistic is unbiased. A classic counterexample is trying to estimate a population's maximum value using the **sample maximum**. Intuitively, a sample can only ever *miss* the true maximum or hit it exactly — it can never overshoot past it — so a systematic downward pull should be expected before any calculation is done at all.

Ví dụ mẫu · The sample maximum systematically underestimates

Let the population be $\{1, 2, \dots, 10\}$, so the true population maximum is 10. Take an SRS of $n = 3$ values without replacement and use the sample maximum as the estimate. There are $\binom{10}{3} = 120$ equally likely samples. For the sample maximum to equal a specific value m, that value m must be in the sample, and the other two values must both come from $\{1, \dots, m-1\}$ — there are $\binom{m-1}{2}$ such samples.

Sample maximum m	3	4	5	6	7	8	9	10
Number of samples $\binom{m-1}{2}$	1	3	6	10	15	21	28	36
Probability (out of 120)	0.008	0.025	0.050	0.083	0.125	0.175	0.233	0.300

The mean of this sampling distribution is

$$\mu_{\text{max}} = \frac{3(1) + 4(3) + 5(6) + 6(10) + 7(15) + 8(21) + 9(28) + 10(36)}{120} = \frac{990}{120} = 8.25.$$

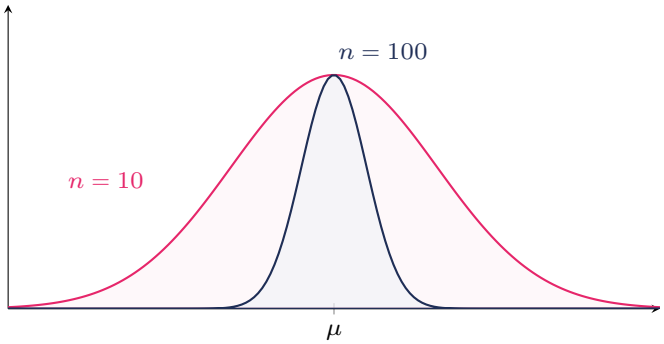
Even though the true population maximum is 10, the sample maximum averages only 8.25 across all possible samples — it **systematically underestimates** the true maximum, because the value 10 itself must be drawn to appear as the sample max, which happens in only 36 of the 120 samples. In general, for an SRS of size k without replacement from $\{1, \dots, N\}$, $E[\text{sample max}] = \dfrac{k(N+1)}{k+1}$, which is always strictly less than N whenever $k < N$; the bias shrinks toward 0 only as $k \to N$ (i.e., only as the "sample" approaches a full census).

A parallel calculation shows the mirror-image effect for the **sample minimum** as an estimator of the population minimum. For the same population $\{1, \dots, 10\}$ and SRS of $n = 3$, the sample minimum equals m only when m is drawn and the other two values come from $\{m+1, \dots, 10\}$ — there are $\binom{10-m}{2}$ such samples for each $m = 1, \dots, 8$:

Sample minimum m	1	2	3	4	5	6	7	8
Number of samples $\binom{10-m}{2}$	36	28	21	15	10	6	3	1

$$\mu_{\text{min}} = \frac{1(36) + 2(28) + 3(21) + 4(15) + 5(10) + 6(6) + 7(3) + 8(1)}{120} = \frac{330}{120} = 2.75.$$

The true population minimum is 1, but the sample minimum averages 2.75 across all possible samples — it **systematically overestimates** the true minimum, the mirror image of the sample maximum's downward bias. (Notice $2.75 = 11 - 8.25$: for the symmetric population $\{1, \dots, 10\}$, the bias of the minimum and the bias of the maximum are exact reflections of each other around the population's center, 5.5.) Both statistics, minimum and maximum, are biased in opposite directions, and neither bias shrinks by collecting a larger sample unless the sample size approaches the full population.



Two sampling distributions of $\bar{x}$ from the same population ($\mu = 50$, $\sigma = 20$), both correctly centered at μ (both unbiased) but with very different spread: $\sigma_{\bar{x}} = 20/\sqrt{10} \approx 6.32$ for $n = 10$ versus $\sigma_{\bar{x}} = 20/\sqrt{100} = 2.00$ for $n = 100$. Bias is about the center; variability is about the spread — the two curves illustrate identical bias (none) with very different variability.

(!) Lỗi thường gặp: Đừng nhầm lẫn "không thiên lệch" (unbiased) với "chính xác / ít biến thiên" (precise). Đây là **hai tính chất khác nhau**. Tăng cỡ mẫu n chỉ **giảm độ biến thiên** (độ trải của phân phối lấy mẫu hẹp lại, ví dụ $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ nhỏ dần) – nó **không** sửa được thiên lệch. Nếu một thống kê vốn đã thiên lệch (như sample maximum ở trên), lấy mẫu lớn hơn chỉ khiến sai số đó lặp lại **chính xác hơn**, chứ phân phối vẫn lệch tâm sai chỗ. Trên đề thi, hãy luôn tự hỏi: câu hỏi đang hỏi về **tâm** (thiên lệch) hay về **độ trải** (biến thiên) của phân phối lấy mẫu – hai câu trả lời không bao giờ được dùng thay thế cho nhau.

5.4 Sampling distribution of a sample proportion

Before this section and the next develop exact formulas, it is worth naming the checklist used throughout the remainder of this unit — the same three conditions, checked in the same order, every time.

The three-step routine: conditions → μ, σ → z

Random: the data must come from a genuine SRS (or a randomized experiment) — without genuine randomization, no probability statement about a sampling distribution is justified. **10% condition:** if sampling without replacement, the sample size must satisfy $n \leq 0.10N$, so that observations behave as close to independent. (Why 10%? Removing observations from a finite population without replacement changes the composition of what remains, so successive draws are not perfectly independent; keeping the sample to at most 10% of the population keeps that dependence small enough to safely ignore.) **Shape condition:** for $\hat{p}$, this means **Large Counts** ($np \geq 10$ and $n(1-p) \geq 10$); for $\bar{x}$ (the following section), this means the

population itself is (approximately) Normal, or n is large enough for the **Central Limit Theorem** to apply (rule of thumb $n \geq 30$). Only once every relevant condition is verified do we compute μ and σ of the sampling distribution, then standardize with z and read the standard Normal table.

VN **GIẢI THÍCH TIẾNG VIỆT**

Trước khi dùng mô hình chuẩn cho phân phối lấy mẫu, luôn kiểm tra đúng thứ tự ba điều kiện: **Random** (SRS hoặc thí nghiệm ngẫu nhiên hoá), **10%** (nếu lấy mẫu không hoàn lại, $n \leq 10\%$ tổng thể), và **điều kiện hình dạng** (Large Counts cho $\hat{p}$; tổng thể chuẩn hoặc $n \geq 30$ cho $\bar{x}$). Quy trình ba bước này không đổi xuyên suốt phần còn lại của Unit: kiểm tra điều kiện → tính μ, σ của phân phối lấy mẫu → chuẩn hoá z và tra bảng chuẩn.

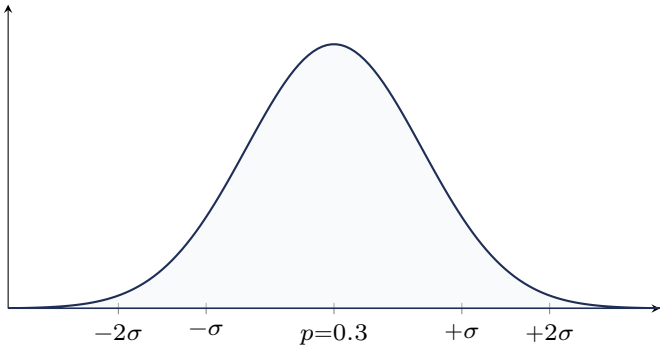
If an SRS of size n is taken from a population with true proportion p: $\mu_{\hat{p}} = p,\ \sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.
The sampling distribution of $\hat{p}$ is approximately Normal when the **Large Counts** condition holds: $np \geq 10$ and $n(1-p) \geq 10$ (and the sample is at most 10% of the population).

Ví dụ mẫu · Worked example

In a population with true proportion $p = 0.3$, an SRS of $n = 100$ is taken. Find $P(\hat{p} > 0.35)$.

Large Counts: $np = 30 \geq 10$, $n(1-p) = 70 \geq 10$ — condition met. $\mu_{\hat{p}} = 0.3$, $\sigma_{\hat{p}} = \sqrt{\dfrac{0.3(0.7)}{100}} = \sqrt{0.0021} \approx 0.0458$.

$$z = \frac{0.35 - 0.3}{0.0458} \approx 1.09 \Rightarrow P(\hat{p} > 0.35) = P(Z > 1.09) \approx 0.1379.$$



Sampling distribution of $\hat{p}$ for $p = 0.3$, $n = 100$: approximately Normal, centered at p, with $\sigma_{\hat{p}} \approx 0.0458$.

This single example already shows the full mechanics: verify conditions, compute $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$, then standardize. Why does the Large Counts condition use the specific threshold 10? It is a rule of thumb, not an exact law: the sampling distribution of $\hat{p}$ is really built from a discrete count (the number of "successes" in the sample), and a continuous Normal curve approximates a discrete distribution well only once there is enough "room" on both sides of the center for the bell shape to form — roughly 10 expected successes and 10 expected failures is the threshold most statisticians have found gives a reasonably accurate approximation in practice. The AP exam expects this threshold to be checked explicitly, not derived from scratch. The next two subsections apply the same three-step routine to a complete four-part write-up, and then to a case where the Normal approximation is not valid at all.

A quick scan across several (p, n) pairs shows how easily Large Counts can pass or fail depending on how close p sits to 0 or 1:

p	n	np	$n(1-p)$	Large Counts?
0.50	20	10	10	met (boundary)
0.10	50	5	45	fails
0.40	30	12	18	met
0.95	100	95	5	fails
0.25	60	15	45	met

Notice that p values close to 0.5 pass Large Counts most easily (since $p(1-p)$ is largest there), while p values near 0 or 1 demand a much larger n before the condition is satisfied. This is exactly the same reason $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ is largest when $p = 0.5$ and shrinks as p moves toward either extreme – the two facts (Large Counts difficulty and maximal spread) share the same underlying cause, the factor $p(1-p)$.

5.4.1 A complete four-part probability check

Every probability question about $\hat{p}$ follows the same four-part routine: (1) find the **center** $\mu_{\hat{p}}$, (2) find the **spread** $\sigma_{\hat{p}}$, (3) verify **shape/conditions** (Random, 10%, Large Counts), and only then (4) **compute the probability** using z-scores.

Ví dụ mẫu · Four-part check for a first-generation student rate

A large university reports that $p = 0.40$ of its (many thousands of) undergraduates are first-generation college students. For an SRS of $n = 150$ students, find $P(\hat{p} < 0.33)$.

(1) Center: $\mu_{\hat{p}} = p = 0.40$.

(2) Spread: $\sigma_{\hat{p}} = \sqrt{\dfrac{0.40(0.60)}{150}} = \sqrt{0.0016} = 0.04$.

(3) Shape/conditions: Random – SRS is given. 10% – 150 is surely at most 10% of the university's undergraduate population. Large Counts – $np = 150(0.40) = 60 \geq 10$ and $n(1-p) = 150(0.60) = 90 \geq 10$. All three conditions are **met**, so the sampling distribution of $\hat{p}$ is approximately Normal.

(4) Probability:

$$z = \frac{0.33 - 0.40}{0.04} = -1.75. \qquad P(\hat{p} < 0.33) = P(Z < -1.75) = 1 - 0.9599 = 0.0401.$$

There is only about a 4.01% chance that a random sample of 150 students would show fewer than 33% first-generation students, even though sampling variability alone could produce such a result.

Ví dụ mẫu · A two-sided probability

A manufacturing line's defect rate is believed to be $p = 0.10$. For an SRS of $n = 400$ units inspected, find $P(0.085 < \hat{p} < 0.115)$.

Conditions: Random (SRS given); 10% (a production run vastly exceeds $4{,}000$ units); Large Counts: $np = 400(0.10) = 40 \geq 10$, $n(1-p) = 400(0.90) = 360 \geq 10$ — **met**.

$$\mu_{\hat{p}} = 0.10, \qquad \sigma_{\hat{p}} = \sqrt{\frac{0.10(0.90)}{400}} = \sqrt{0.000225} = 0.015.$$

$$z_{\text{low}} = \frac{0.085 - 0.10}{0.015} = -1.00, \qquad z_{\text{high}} = \frac{0.115 - 0.10}{0.015} = 1.00.$$

$$P(0.085 < \hat{p} < 0.115) = P(-1.00 < Z < 1.00) = 0.8413 - (1 - 0.8413) = 0.6826.$$

This is the familiar "within one standard deviation" figure from the 68–95–99.7 rule (Unit 1) — the same empirical rule applies to *any* approximately Normal distribution, including the sampling distribution of a statistic.

5.4.2 When the Large Counts condition fails

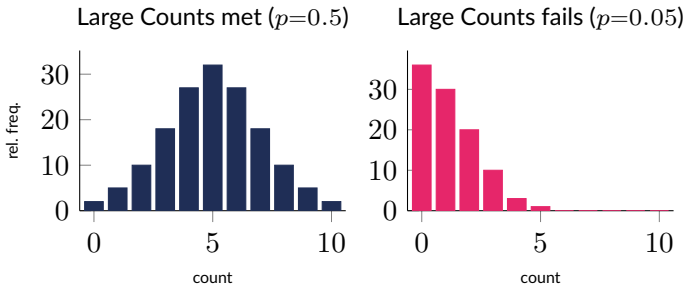
Not every proportion problem allows a Normal approximation. When the true proportion p is close to 0 or 1, and n is not large enough to compensate, the sampling distribution of $\hat{p}$ can be badly skewed.

Ví dụ mẫu · A rare condition – Large Counts fails

A rare genetic condition affects $p = 0.015$ of newborns nationally. For an SRS of $n = 400$ newborns, can the sampling distribution of $\hat{p}$ be treated as approximately Normal?

Random: SRS given. **10%:** any reasonable newborn population vastly exceeds $4{,}000$, so 10% holds easily. **Large Counts:** $np = 400(0.015) = 6 < 10$ — this already **fails**, even though $n(1-p) = 400(0.985) = 394 \geq 10$ easily. Since *both* parts of Large Counts must hold, the condition is not satisfied.

Because the expected number of affected newborns in the sample is only 6, the count (and hence $\hat{p}$) behaves like a strongly right-skewed discrete distribution: most samples will show a small handful of cases (0, 1, 2, ...), a symmetric bell curve would badly misrepresent this skew, and it could even assign meaningful probability to nonsensical outcomes near $\hat{p} < 0$. Normal-based z-score methods are therefore **not valid** here; a probability question about this sample proportion would require different tools (e.g., the exact binomial distribution or simulation), not covered by the Normal model of this unit. How large would n need to be for Large Counts to hold with this same $p = 0.015$? We would need $np \geq 10$, i.e., $n \geq 10/0.015 \approx 666.7$, so at least $n = 667$ newborns — well over the $n = 400$ actually available in this example, and nearly 20 times the minimum $n \approx 34$ that would suffice for a proportion closer to the center, such as $p = 0.3$. The further p sits from 0.5, the larger a sample must be before a Normal approximation becomes trustworthy.



Illustrative comparison of two sampling-distribution shapes for a count out of $n = 20$ trials: when Large Counts is met (left, e.g. $p = 0.5$), the shape is roughly symmetric and bell-shaped; when Large Counts fails (right, p close to 0), the shape is noticeably right-skewed and clustered near 0 — a Normal curve would poorly describe it.

VN **GIẢI THÍCH TIẾNG VIỆT**

Với $\hat{p}$: kiểm tra đủ **ba điều kiện** Random – 10% – Large Counts trước khi dùng mô hình chuẩn. Khi thoả, $\mu_{\hat{p}} = p$ (không chệch) và $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, rồi chuẩn hoá z và tra bảng như bình thường. Khi p rất gần 0 hoặc 1 và n không đủ lớn để bù lại (Large Counts thất bại), phân phối của $\hat{p}$ bị lệch mạnh và **không được** áp dụng mô hình chuẩn – đây là lỗi rất hay gặp khi học sinh

quên kiểm tra điều kiện trước khi tính z.

5.5 Sampling distribution of a sample mean

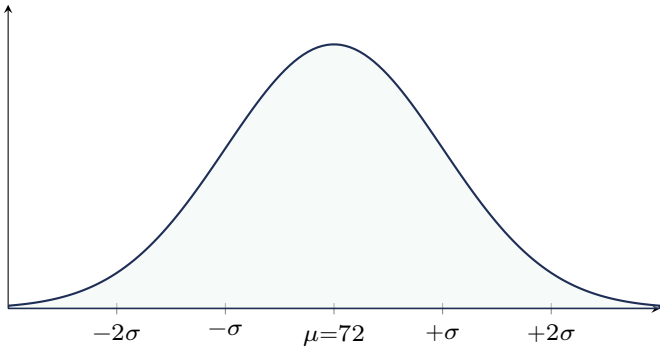
The same three-step routine used for $\hat{p}$ applies to $\bar{x}$, with one important difference in the shape condition: instead of Large Counts, we now ask whether the *population itself* is Normal, or whether n is large enough for the Central Limit Theorem to take over.

For an SRS of size n from a population with mean μ and standard deviation σ: $\mu_{\bar{x}} = \mu$, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$. **Central Limit Theorem:** as n grows large, the sampling distribution of $\bar{x}$ becomes approximately Normal *regardless of the population's shape* (a common rule of thumb is $n \geq 30$). If the population itself is Normal, $\bar{x}$ is exactly Normal for any n.

Ví dụ mẫu · Worked example

Adult resting heart rate has $\mu = 72$ bpm, $\sigma = 9$ bpm (population shape roughly Normal). For an SRS of $n = 36$, find $P(\bar{x} > 75)$.

$\sigma_{\bar{x}} = \dfrac{9}{\sqrt{36}} = \dfrac{9}{6} = 1.5$. $z = \dfrac{75 - 72}{1.5} = 2.0 \Rightarrow P(\bar{x} > 75) = P(Z > 2.0) \approx 0.0228$.



Sampling distribution of $\bar{x}$ for SRS of $n = 36$ resting heart rates ($\mu = 72$, $\sigma = 9$): approximately Normal, centered at 72, with $\sigma_{\bar{x}} = 1.5$.

The heart-rate example above worked because the population itself was stated to be roughly Normal. Because $\sigma_{\bar{x}}$ (like $\sigma_{\hat{p}}$) measures how much a *statistic* varies from sample to sample, rather than how much a single individual observation varies, it is commonly called the **standard error** of the statistic — terminology used throughout the confidence-interval and significance-testing units later in this course. The next three subsections examine what happens to the sampling distribution of $\bar{x}$ under three different combinations of population shape and sample size: large n from a skewed population, small n from a Normal population, and (as a caution) small n from a skewed population.

(!) Lỗi thường gặp: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ giảm khi n tăng, nhưng **không** bằng 0 và không thể tính nếu không biết σ (khi đó phải dùng s và phân phối t — sẽ học ở Unit 7). Đừng nhầm $\sigma_{\bar{x}}$ (độ lệch chuẩn của $\bar{x}$) với σ (độ lệch chuẩn của một quan sát đơn lẻ).

5.5.1 The Central Limit Theorem on a skewed population

The power of the CLT is that it says nothing about the shape of the *population* — only about the sampling distribution of $\bar{x}$, once n is large enough.

A right-skewed population of household incomes – most households have low-to-moderate income, with a long tail of high earners. For this illustration, $\mu = \$65{,}000$ and $\sigma = \$40{,}000$. The population itself does not need to be Normal for the Central Limit Theorem to apply to $\bar{x}$.

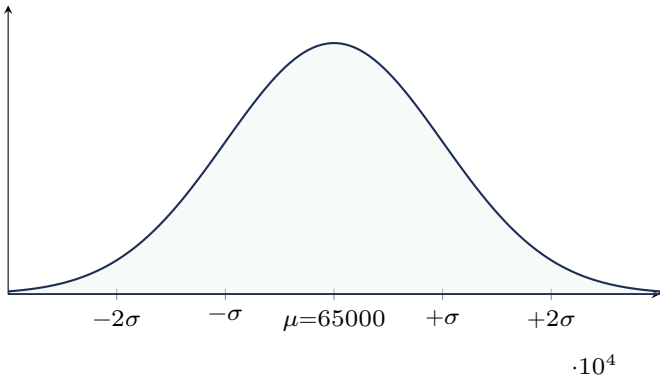
Ví dụ mẫu · CLT applied to a skewed population

Household incomes in a region are strongly right-skewed with $\mu = \$65{,}000$ and $\sigma = \$40{,}000$. For an SRS of $n = 100$ households, find $P(\bar{x} > \$70{,}000)$.

Conditions: Random (SRS given); 10% (any real region has far more than $1{,}000$ households); population shape is skewed, *not* Normal, so we cannot claim $\bar{x}$ is exactly Normal for small n — but $n = 100 \geq 30$, so by the **Central Limit Theorem** the sampling distribution of $\bar{x}$ is approximately Normal *regardless of the population's skew*.

$$\mu_{\bar{x}} = 65000, \qquad \sigma_{\bar{x}} = \frac{40000}{\sqrt{100}} = \frac{40000}{10} = 4000.$$

$$z = \frac{70000 - 65000}{4000} = 1.25. \qquad P(\bar{x} > 70000) = P(Z > 1.25) = 1 - 0.8944 = 0.1056.$$



The sampling distribution of $\bar{x}$ for SRS of $n = 100$ household incomes: already approximately Normal in shape by the CLT, centered at the same mean $\mu = \$65{,}000$, but with much smaller spread ($\sigma_{\bar{x}} = \$4{,}000$ versus the population's own $\sigma = \$40{,}000$).

5.5.2 A small sample from a Normal population

When the population itself is already (approximately) Normal, the CLT is not even needed — $\bar{x}$ is exactly Normal for *any* sample size, even a very small one.

Ví dụ mẫu · Small n, but the population is already Normal

A machine fills specialty coffee bags with weights that are Normally distributed, $\mu = 500$ g, $\sigma = 15$ g. For a small SRS of only $n = 9$ bags, find $P(\bar{x} < 493 \text{ g})$.

Conditions: Random (SRS given); 10% (a production run vastly exceeds 90 bags); the *population* is stated to be Normal, so the sampling distribution of $\bar{x}$ is exactly Normal for *any* n, including this small $n = 9$ — no need to invoke the CLT or check $n \geq 30$.

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{9}} = \frac{15}{3} = 5. \qquad z = \frac{493 - 500}{5} = -1.4.$$

$$P(\bar{x} < 493) = P(Z < -1.4) = 1 - 0.9192 = 0.0808.$$

The key phrase to notice in this problem is "weights that are Normally distributed" — that single fact does all the work of satisfying the shape condition, regardless of how small n is.

5.5.3 A caution: small samples from a skewed population

The two examples above work because in each case at least one condition (Normal population, or $n \geq 30$) held. When *neither* holds, no valid probability calculation is possible with the tools of this unit — recognizing this situation, and refusing to compute a number anyway, is itself a tested skill on the AP exam.

Ví dụ mẫu · Neither condition holds – do not compute a probability

Customer wait times at a help desk are strongly right-skewed (most calls resolve quickly, but a few take very long), with $\mu = 4$ minutes and $\sigma = 6$ minutes. For a small SRS of only $n = 4$ calls, a student wants $P(\bar{x} > 7)$.

Conditions: the population is explicitly *not* Normal (strongly skewed), and $n = 4$ is nowhere near the $n \geq 30$ threshold needed for the CLT to overcome that skew. **Neither** condition for treating $\bar{x}$ as approximately Normal is satisfied.

If a student naively plugged into the Normal formula anyway: $\sigma_{\bar{x}} = 6/\sqrt{4} = 3$, $z = (7-4)/3 = 1.00$, suggesting $P(\bar{x} > 7) \approx 1 - 0.8413 = 0.1587$ — but this number is **not trustworthy**. With only 4 skewed observations averaged together, the true sampling distribution of $\bar{x}$ is itself still visibly skewed, and z-scores from the standard Normal table do not apply. The correct response to such a problem is to state that the probability **cannot be determined** with the methods of this unit, not to report a numerical answer. (In practice, statisticians facing this situation would instead use simulation-based methods — such as bootstrap resampling — that do not require the Normal-shape assumption at all; those methods are outside the scope of this unit, but the conditions check above is exactly how a student recognizes when they would be needed.)

VN **GIẢI THÍCH TIẾNG VIỆT**

Với $\bar{x}$: kiểm tra Random – 10% – (tổng thể chuẩn HOẶC n đủ lớn để áp dụng CLT, thường lấy mốc $n \geq 30$). Nếu tổng thể đã cho là chuẩn, $\bar{x}$ luôn chuẩn với **mọi** n, kể cả n rất nhỏ. Nếu tổng thể lệch nhưng $n \geq 30$, CLT vẫn đảm bảo $\bar{x}$ xấp xỉ chuẩn. Nhưng nếu tổng thể lệch **và** n nhỏ, KHÔNG được giả định $\bar{x}$ xấp xỉ chuẩn – đừng cố tính z-score trong trường hợp này, vì kết quả sẽ không đáng tin cậy.

5.6 How sample size affects the sampling distribution

Both $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$ shrink as n grows — but only through the *square root* of n, not n itself. This has a very concrete practical consequence for anyone designing a study or a survey: precision does not come cheaply, and doubling the effort spent collecting data does not double the precision gained.

Because $\sigma_{\bar{x}}$ and $\sigma_{\hat{p}}$ are both proportional to $1/\sqrt{n}$, cutting a sampling distribution's standard deviation **in half** requires **quadrupling** n (multiplying by 4, not 2). More generally, to shrink the standard deviation by a factor of k, the sample size must increase by a factor of k^2.

Ví dụ mẫu · How much bigger must n be to halve the spread?

A poll currently uses $n = 400$ respondents to estimate a proportion believed to be $p = 0.5$. Its standard deviation is

$$\sigma_{\hat{p}} = \sqrt{\frac{0.5(0.5)}{400}} = \sqrt{0.000625} = 0.025.$$

How large would the new sample need to be to cut this standard deviation exactly in half, to 0.0125?

Set $\sqrt{\dfrac{0.25}{n_{\text{new}}}} = 0.0125$. Squaring both sides: $\dfrac{0.25}{n_{\text{new}}} = 0.00015625$, so $n_{\text{new}} = \dfrac{0.25}{0.00015625} = 1600$.

Check: $\sigma_{\hat{p}} = \sqrt{0.25/1600} = \sqrt{0.00015625} = 0.0125$ — exactly half of the original 0.025. Increasing n from 400 to 1600 is exactly a factor of 4 (i.e., 2^2), confirming that halving $\sigma_{\hat{p}}$ requires **quadrupling** the sample size, not merely doubling it.

The same $1/\sqrt{n}$ pattern holds at every scale, not just for the single jump worked out above:

n	25	100	400	1600
$\sigma_{\hat{p}}$ (with $p = 0.5$)	0.1000	0.0500	0.0250	0.0125

Each time n quadruples ($25 \to 100 \to 400 \to 1600$), $\sigma_{\hat{p}}$ is cut exactly in half — the same pattern repeats at every step, not just the one step worked out above. This is also why collecting more data has **diminishing returns**: the jump from $n = 25$ to $n = 100$ (a difference of only 75 observations) cuts the standard deviation in half, but so does the much larger jump from $n = 400$ to $n = 1600$ (a difference of 1200 observations) — each additional halving of the spread costs four times as much data as the halving before it. This is precisely why researchers designing real studies must balance the cost of collecting additional data against the shrinking benefit each additional observation provides.

VN **GIẢI THÍCH TIẾNG VIỆT**

Vì $\sigma_{\bar{x}}$ và $\sigma_{\hat{p}}$ tỉ lệ với $1/\sqrt{n}$ (chứ không phải $1/n$), muốn giảm độ lệch chuẩn của phân phối lấy mẫu xuống còn **một nửa**, ta phải tăng n lên gấp **bốn lần**, không phải gấp đôi. Đây là lỗi rất phổ biến: học sinh thường nghĩ "muốn độ chính xác gấp đôi thì lấy mẫu gấp đôi" – nhưng thực tế phải lấy mẫu gấp **bốn lần** (tổng quát: muốn giảm độ lệch chuẩn đi k lần, phải tăng n lên k^2 lần).

5.7 Combining sampling distributions: differences between two samples

The ideas above extend naturally to comparing *two* independent samples — the foundation for two-sample inference in later units (Unit 6–8). Every two-sample confidence interval and significance test built later in this course — comparing two proportions, or comparing two means — is constructed directly from the sampling distribution developed in this section, so the formulas below are worth committing to memory now rather than re-deriving later.

For two **independent** random samples: the sampling distribution of a **difference of two sample proportions** $\hat{p}_1 - \hat{p}_2$ has

$$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2, \qquad \sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},$$

and the sampling distribution of a **difference of two sample means** $\bar{x}_1 - \bar{x}_2$ has

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2, \qquad \sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

In both cases the mean of the difference is the **difference of the parameters**, and because the two samples are independent, the **variances add** (never the standard deviations directly) before the square root is taken.

The requirement of **independence** between the two samples is essential and must be checked, not assumed: the variance-addition formula above is valid only when knowing the outcome of Sample 1 tells us nothing about Sample 2 — for example, two separately drawn SRSs from two different populations, or two randomly assigned groups in an experiment. It would *not* apply to before-and-after measurements taken on the *same* individuals, since those two sets of measurements are related to each other by design, not independent.

Why do independent variances *add*, even though the statistic itself is a *difference*? This follows from a general property of independent random variables: for any two independent quantities X and Y, $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y)$. Subtraction affects the *mean* ($E[X-Y] = E[X] - E[Y]$), but variance is always additive under independence, whether the two quantities are added or subtracted — a negative sign squares away when variance is computed. Applying this general rule directly to $\hat{p}_1$ and $\hat{p}_2$ (or $\bar{x}_1$ and $\bar{x}_2$) produces exactly the variance-addition formulas above.

VN **GIẢI THÍCH TIẾNG VIỆT**

Khi kết hợp hai mẫu **độc lập**, tâm của phân phối lấy mẫu của **hiệu** hai thống kê luôn là **hiệu hai tham số thật** ($p_1 - p_2$ hoặc $\mu_1 - \mu_2$) – không thiên lệch. Nhưng độ lệch chuẩn thì KHÔNG được trừ hay cộng trực tiếp – ta phải **cộng phương sai** của hai mẫu lại rồi mới lấy căn bậc hai. Ý tưởng này sẽ được dùng lại xuyên suốt các Unit suy diễn hai mẫu (so sánh hai tỉ lệ, so sánh hai trung bình) ở các chương sau. Ghi nhớ: hai mẫu phải **độc lập** với nhau (không phải cùng một nhóm đối tượng đo hai lần) thì công thức cộng phương sai này mới có giá trị.

Ví dụ mẫu · Comparing two independent approval rates

Company A reports $p_1 = 0.90$ approval among an SRS of $n_1 = 100$ employees. Company B reports $p_2 = 0.64$ approval among an independent SRS of $n_2 = 144$ employees. Find $P(\hat{p}_1 - \hat{p}_2 > 0.20)$.

Conditions: both samples random and independent (given); each sample is surely at most 10% of its company; Large Counts for Company A: $n_1p_1 = 90 \geq 10$, $n_1(1-p_1) = 10 \geq 10$; Large Counts for Company B: $n_2p_2 = 92.16 \geq 10$, $n_2(1-p_2) = 51.84 \geq 10$ — all **met**.

$$\mu_{\hat{p}_1-\hat{p}_2} = 0.90 - 0.64 = 0.26.$$

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{0.90(0.10)}{100} + \frac{0.64(0.36)}{144}} = \sqrt{0.0009 + 0.0016} = \sqrt{0.0025} = 0.05.$$

$$z = \frac{0.20 - 0.26}{0.05} = -1.20. \qquad P(\hat{p}_1 - \hat{p}_2 > 0.20) = P(Z > -1.20) = P(Z < 1.20) = 0.8849.$$

There is about an 88.49% chance that a pair of independent samples like these would show Company A's approval rate exceeding Company B's by more than 0.20, consistent with the true gap of 0.26. Note that even though $n_1 \neq n_2$ and $p_1 \neq p_2$, the same three-step routine applies without modification: verify conditions for *each* sample separately, compute μ and σ of the difference, then standardize.

The same logic applies directly to a difference of two independent *means*, replacing $p(1-p)/n$ with σ^2/n inside the sum of variances.

Ví dụ mẫu · Comparing two independent means

Line 1 produces bolts with lengths approximately Normal, $\mu_1 = 50.00$ mm, $\sigma_1 = 0.3$ mm, from an SRS of $n_1 = 100$ bolts. Line 2 produces bolts with lengths approximately Normal, $\mu_2 = 50.05$ mm, $\sigma_2 = 0.4$ mm, from an independent SRS of $n_2 = 100$ bolts. Find $P(\bar{x}_1 - \bar{x}_2 > 0)$ — the chance that Line 1's *sample* mean comes out higher than Line 2's, even though Line 1's *true* mean is lower.

Conditions: both samples random and independent (given); both populations stated to be approximately Normal, so the Normal condition holds regardless of n.

$$\mu_{\bar{x}_1 - \bar{x}_2} = 50.00 - 50.05 = -0.05.$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{0.3^2}{100} + \frac{0.4^2}{100}} = \sqrt{0.0009 + 0.0016} = \sqrt{0.0025} = 0.05.$$

$$z = \frac{0 - (-0.05)}{0.05} = 1.00. \qquad P(\bar{x}_1 - \bar{x}_2 > 0) = P(Z > 1.00) = 1 - 0.8413 = 0.1587.$$

Even though Line 1's bolts are, on average, slightly shorter than Line 2's, sampling variability alone gives roughly a 15.87% chance that a particular pair of samples would show the opposite pattern — a reminder of why a single pair of sample means is rarely enough, by itself, to prove one population's true mean exceeds another's. This exact kind of reasoning, formalized with a decision rule for how small a probability must be before it counts as convincing evidence, becomes the significance test for two independent means in a later unit.

Summary of this unit's core formulas:

$$\mu_{\hat{p}} = p, \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}; \qquad \mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}};$$

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2, \quad \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}; \qquad \mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2, \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

In every case the same three-step routine applies: verify conditions (Random, 10%, Large Counts or Normal/$n \geq 30$, checked separately for each sample when there are two) → compute the mean and standard deviation of the sampling distribution → standardize with z and read the standard Normal table.

Statistic	Conditions	Center and spread
$\hat{p}$	Random; 10%; Large Counts ($np \geq 10$, $n(1-p) \geq 10$)	$\mu_{\hat{p}} = p, \quad \sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$
$\bar{x}$	Random; 10%; population Normal, or $n \geq 30$ (CLT)	$\mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$
$\hat{p}_1 - \hat{p}_2$	Random & independent; 10% for each; Large Counts for *each* sample separately	$\mu = p_1 - p_2, \quad \sigma = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$
$\bar{x}_1 - \bar{x}_2$	Random & independent; 10% for each; both populations Normal, or both $n_1, n_2 \geq 30$	$\mu = \mu_1 - \mu_2, \quad \sigma = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

A handful of pairs of ideas from this unit are especially easy to mix up under exam pressure; it is worth keeping them straight explicitly.

Often confused	First quantity	Second quantity
Variability of an individual vs. of a statistic	σ (a single observation)	$\sigma_{\bar{x}} = \sigma/\sqrt{n}$, or $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$
Bias vs. variability	the *center* of a sampling distribution	the *spread* of a sampling distribution
Large Counts vs. CLT	shape condition for $\hat{p}$ ($np \geq 10$, $n(1-p) \geq 10$)	shape condition for $\bar{x}$ (Normal population, or $n \geq 30$)
Doubling n vs. quadrupling n	doubles $n \Rightarrow$ divides $\sigma_{\text{statistic}}$ by $\sqrt{2}$	quadruples $n \Rightarrow$ divides $\sigma_{\text{statistic}}$ by 2

5.8 Why sampling distributions matter beyond this unit

Every idea developed in this chapter reappears, essentially unchanged, in the four units that follow. A **confidence interval** is nothing more than a statistic plus-or-minus a multiple of its standard error, statistic $\pm\, z^* \cdot \sigma_{\text{statistic}}$ (or $t^* \cdot SE_{\text{statistic}}$ once σ must be estimated by s), built directly from the center and spread of a sampling distribution exactly like those developed here. A **significance test** asks precisely the kind of question already practiced throughout this unit – "if the null hypothesis were true, how unusual would this sample's statistic be?" – by computing a z (or t) statistic from the sampling distribution and comparing it against a standard reference distribution.

Three properties of a sampling distribution decide whether – and how – a given inference procedure may be used: its **shape** (is a Normal or t model justified, via Large Counts, the CLT, or a stated Normal population?), its **center** (is the statistic unbiased, so the sampling distribution is centered at the true parameter?), and its **spread** (how is $\sigma_{\text{statistic}}$, or its estimate, computed for this particular statistic?). Every inference procedure introduced in Units 6 through 8 begins by answering exactly these three questions for a specific statistic: $\hat{p}$, $\bar{x}$, $\hat{p}_1 - \hat{p}_2$, $\bar{x}_1 - \bar{x}_2$, and others introduced later.

VN **GIẢI THÍCH TIẾNG VIỆT**

Toàn bộ nội dung Unit này sẽ được dùng lại gần như nguyên vẹn trong các Unit suy diễn (Unit 6–8): **khoảng tin cậy (confidence interval)** chỉ là "thống kê $\pm$ một hệ số nhân $\times$ sai số chuẩn (standard error)" – lấy trực tiếp từ tâm và độ trải của phân phối lấy mẫu; **kiểm định giả thuyết (significance test)** chỉ là hỏi "nếu giả thuyết gốc đúng, kết quả mẫu này bất thường đến mức

nào?" bằng cách chuẩn hoá z (hoặc t), đúng như đã luyện tập xuyên suốt Unit này. Ba câu hỏi luôn lặp lại ở mọi Unit suy diễn phía sau: hình dạng (mô hình chuẩn có hợp lệ không?), tâm (thống kê có không thiên lệch không?), và độ trải (sai số chuẩn tính thế nào?).

5.9 Practice problems

The items below span every idea developed in this unit — from hand-built sampling distributions and the bias/variability distinction, through the Large Counts and Central Limit Theorem conditions, to sample-size effects and combining two independent samples.

Câu 1

A state election board wants to know the true proportion of all registered voters who support a proposed law. It surveys a random sample of 800 registered voters; 512 say they support it. Identify the population, the parameter, the sample, and the statistic in this context.

Lời giải chi tiết

Population: all registered voters in the state. **Parameter:** the true (unknown) proportion of *all* registered voters who support the law, denoted p. **Sample:** the 800 registered voters who were actually surveyed. **Statistic:** the sample proportion $\hat{p} = 512/800 = 0.64$, computed from the sample and used as a point estimate of p. This fixed-versus-random distinction — one unchanging population parameter versus one sample's statistic, which would come out differently with a different sample — underlies every inference procedure introduced later in this course.

Câu 2

Which statement correctly describes sampling variability?

(A) Sampling variability is an error that occurs only when a biased sampling method is used.

(B) Sampling variability is the natural tendency of a statistic's value to differ from sample to sample, even under the exact same correct random sampling method.

(C) Sampling variability is eliminated once the sample size is at least 30.

(D) Sampling variability describes the difference between two separate populations.

Lời giải chi tiết

Sampling variability is the natural, unavoidable tendency of a statistic to take different values across different random samples of the same size, drawn by the exact same correct method — it is not a mistake, and it shrinks with n but never fully disappears. **(B)**.

Câu 3

Consider the small population $\{1, 3, 5, 7, 9\}$. List every possible SRS of size $n = 2$ (without replacement), compute $\bar{x}$ for each, display the resulting sampling distribution of $\bar{x}$ in a table, and verify that its mean equals the population mean.

Lời giải chi tiết

Population mean: $\mu = \dfrac{1+3+5+7+9}{5} = \dfrac{25}{5} = 5$. There are $\binom{5}{2} = 10$ equally likely samples:

Sample	$\{1,3\}$	$\{1,5\}$	$\{1,7\}$	$\{1,9\}$	$\{3,5\}$	$\{3,7\}$	$\{3,9\}$	$\{5,7\}$	$\{5,9\}$	$\{7,9\}$
$\bar{x}$	2	3	4	5	4	5	6	6	7	8

Sampling distribution of $\bar{x}$ (each sample has probability $1/10$):

$\bar{x}$	2	3	4	5	6	7	8
Probability	$1/10$	$1/10$	$2/10$	$2/10$	$2/10$	$1/10$	$1/10$

Mean of the sampling distribution: $\mu_{\bar{x}} = \dfrac{2+3+4+4+5+5+6+6+7+8}{10} = \dfrac{50}{10} = 5$, which equals the population mean $\mu = 5$ exactly — confirming $\bar{x}$ is unbiased. Notice, too, that the sampling distribution's values (2 through 8) already cluster more tightly around 5 than a uniform spread across the full range would suggest, since the middle value $\bar{x} = 5$ occurs most often (2 of the 10 samples).

Câu 4

Consider the small population $\{10, 20, 30\}$. List every possible SRS of size $n = 2$ (without replacement), compute $\bar{x}$ for each, and verify that the mean of the resulting sampling distribution equals the population mean.

Lời giải chi tiết

Population mean: $\mu = \dfrac{10+20+30}{3} = \dfrac{60}{3} = 20$. There are $\binom{3}{2} = 3$ equally likely samples: $\{10, 20\} \to \bar{x} = 15$; $\{10, 30\} \to \bar{x} = 20$; $\{20, 30\} \to \bar{x} = 25$. Each has probability $1/3$. Mean of the sampling distribution: $\dfrac{15+20+25}{3} = \dfrac{60}{3} = 20$, exactly equal to the population mean $\mu = 20$ — confirming $\bar{x}$ is unbiased even for this very small population, where only 3 distinct samples exist at all.

Câu 5

Which of the following statistics, computed from repeated SRSs of a population, is an unbiased estimator of the population mean μ?

(A) The sample mean $\bar{x}$

(B) The sample maximum

(C) The sample range (maximum − minimum)

(D) The value of the very first observation drawn in the sample

Lời giải chi tiết

$\bar{x}$ is unbiased: $\mu_{\bar{x}} = \mu$ for any sample size n. The sample maximum systematically underestimates the population maximum (as shown in this unit); the sample range and a single first observation are not generally unbiased estimators of μ either, since neither one's long-run average across repeated samples equals μ in general. **(A)**.

Câu 6

Two methods for estimating the true weight of a reference object (known to weigh exactly 50.0 kg) are compared by repeating each measurement many times. Method 1 (a calibrated digital scale) gives readings that average 49.8 kg across repetitions, tightly clustered between 49.6 and 50.0 kg. Method 2 (an uncalibrated spring scale) gives readings that average *exactly* 50.0 kg across repetitions, but individual readings vary widely, from 40 kg to 60 kg. Compare the bias and variability of the two methods.

Lời giải chi tiết

Method 1 has a small **bias** (its long-run average, 49.8 kg, is 0.2 kg below the true 50.0 kg) but **low variability** (readings are tightly clustered — precise and consistent). **Method 2** is **unbiased** (its long-run average equals the true value exactly) but has **high variability** (any single reading could be far off, making it unreliable in a single use). This illustrates that bias and variability are independent properties: an estimator can be almost-unbiased-but-imprecise, or slightly-biased-but-highly-consistent, and neither is automatically "better" without specifying which property matters more. In many practical settings a small, consistent bias (Method 1) is actually preferred over large unpredictable swings (Method 2), since the bias can potentially be corrected once it is known, while high variability cannot.

Câu 7

Which statement correctly compares two SRS's of sizes $n = 25$ and $n = 400$ from the same population, both used to estimate μ with $\bar{x}$?

(A) The $n = 400$ sample gives a biased estimate but $n = 25$ does not

(B) Both $\bar{x}$'s are unbiased for μ, but $\sigma_{\bar{x}}$ is smaller for $n = 400$, so its sampling distribution is less variable

(C) $\sigma_{\bar{x}}$ does not depend on n

(D) The $n = 25$ sample is unbiased only if the population is Normal

Lời giải chi tiết

Both are SRSs, so both give an unbiased estimator ($\mu_{\bar{x}} = \mu$ regardless of n). But $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is smaller for the larger sample ($n = 400$ shrinks $\sigma_{\bar{x}}$ by a factor of $\sqrt{400/25} = 4$ compared to $n = 25$), so the larger sample's statistic is less variable, not less biased. **(B)**.

Câu 8

A population has $p = 0.6$. For an SRS of $n = 20$, is the Large Counts condition met, and why does this matter?

(A) Yes; $np = 12, n(1-p) = 8$, both ≥ 10 so Normal approx. is valid

(B) No; $n(1-p) = 8 < 10$, so the sampling distribution of $\hat{p}$ may be skewed and Normal approximation is unreliable

(C) Yes, because $n \geq 10$

(D) No, because $p > 0.5$

Lời giải chi tiết

$np = 20(0.6) = 12 \geq 10$ but $n(1-p) = 20(0.4) = 8 < 10$. The condition **fails**, so we should not treat the sampling distribution of $\hat{p}$ as approximately Normal here. **(B)**.

Câu 9

A population has true proportion $p = 0.02$. For a proposed SRS of $n = 400$, is the Large Counts condition satisfied?

(A) Yes, since $n = 400$ is a large sample size

(B) Yes, since $n(1-p) = 392 \geq 10$

(C) No, since $np = 400(0.02) = 8 < 10$, even though $n(1-p) = 392 \geq 10$

(D) No, since neither part of the condition is satisfied

Lời giải chi tiết

$np = 400(0.02) = 8 < 10$ — this alone is enough to fail Large Counts, even though $n(1-p) = 392 \geq 10$ comfortably. Both parts must hold; one failing part fails the whole condition, so a Normal model should not be used to compute probabilities for $\hat{p}$ here. **(C)**.

Câu 10

A population has true proportion $p = 0.85$. For a proposed SRS of $n = 50$, is the Large Counts condition satisfied?

(A) Yes, since $np = 42.5 \geq 10$

(B) No, since $n(1-p) = 50(0.15) = 7.5 < 10$, even though $np = 42.5 \geq 10$

(C) Yes, since $p > 0.5$

(D) No, since $np < 10$

Lời giải chi tiết

$np = 50(0.85) = 42.5 \geq 10$ (✓), but $n(1-p) = 50(0.15) = 7.5 < 10$ **(fails)**. Both parts must be checked separately, and this sample fails on the second part — with p close to 1, it is the *failures*, not the successes, that become too rare for a Normal approximation. **(B)**.

Câu 11

A population has true proportion $p = 0.50$. For a proposed SRS of $n = 15$, check whether the Large Counts condition is satisfied, and explain what this means for the shape of the sampling distribution of $\hat{p}$.

Lời giải chi tiết

$np = 15(0.50) = 7.5 < 10$ and $n(1-p) = 15(0.50) = 7.5 < 10$ — **both** parts fail (not just one). Even though $p = 0.5$ is the "most favorable" value of p for Large Counts (it maximizes $p(1-p)$), the sample size $n = 15$ is simply too small to guarantee a Normal-shaped sampling distribution for $\hat{p}$. With so few observations, $\hat{p}$ can only take a small number of discrete values $(0/15, 1/15, \ldots, 15/15)$, and its distribution is better approximated directly by the binomial distribution than by a continuous Normal curve.

Câu 12

Cereal box weights are Normal with $\mu = 68$ g, $\sigma = 3$ g. For an SRS of $n = 36$ boxes, find $P(\bar{x} < 67)$.

(A) 0.0228

(B) 0.3707

(C) 0.4522

(D) 0.9772

Lời giải chi tiết

$\sigma_{\bar{x}} = 3/\sqrt{36} = 0.5$. $z = (67-68)/0.5 = -2.0$. $P(\bar{x} < 67) = P(Z < -2.0) \approx 0.0228$. **(A).**

Câu 13

Battery lifetimes in a large shipment are approximately Normal with $\mu = 100$ hours, $\sigma = 20$ hours. For an SRS of $n = 25$ batteries, find $P(\bar{x} < 98)$.

Lời giải chi tiết

Conditions: SRS given (Random); $n = 25$ surely $\leq 10\%$ of the shipment; population stated to be Normal, so the sampling distribution of $\bar{x}$ is approximately Normal regardless of sample size — **met**.

$$\sigma_{\bar{x}} = \frac{20}{\sqrt{25}} = \frac{20}{5} = 4. \qquad z = \frac{98-100}{4} = -0.5.$$

$$P(\bar{x} < 98) = P(Z < -0.5) = 1 - 0.6915 = 0.3085.$$

There is roughly a 30.85% chance that a random sample of 25 batteries would show an average lifetime below 98 hours, purely from sampling variability around the true mean of 100 hours.

Câu 14

A true population proportion is $p = 0.25$. For an SRS of $n = 300$, find $P(0.20 < \hat{p} < 0.30)$.

Lời giải chi tiết

Conditions: SRS given; $n = 300$ surely $\leq 10\%$ of the population; Large Counts: $np = 75 \geq 10$, $n(1-p) = 225 \geq 10$ — **met**.

$$\sigma_{\hat{p}} = \sqrt{\frac{0.25(0.75)}{300}} = \sqrt{0.000625} = 0.025.$$

$$z_{\text{low}} = \frac{0.20-0.25}{0.025} = -2.00, \qquad z_{\text{high}} = \frac{0.30-0.25}{0.025} = 2.00.$$

$$P(0.20 < \hat{p} < 0.30) = P(-2.00 < Z < 2.00) = 0.9772 - (1 - 0.9772) = 0.9544.$$

This matches the familiar "within two standard deviations" figure ($\approx 95\%$) from the 68–95–99.7 rule, applied here to the sampling distribution of $\hat{p}$ rather than to individual observations.

Câu 15

A machine fills bottles with a Normally distributed volume, $\mu = 355$ mL, $\sigma = 4$ mL. For a small SRS of only $n = 4$ bottles, find $P(\bar{x} < 353)$.

Lời giải chi tiết

Conditions: SRS given; $n = 4$ surely $\leq 10\%$ of the production run; population stated to be Normal, so $\bar{x}$ is exactly Normal for any n, including this very small $n = 4$.

$$\sigma_{\bar{x}} = \frac{4}{\sqrt{4}} = \frac{4}{2} = 2. \qquad z = \frac{353-355}{2} = -1.00.$$

$$P(\bar{x} < 353) = P(Z < -1.00) = 1 - 0.8413 = 0.1587.$$

Even with only 4 bottles in the sample, the calculation is fully valid, because the *population* of individual bottle volumes was already stated to be Normal — no CLT or large-n argument is needed here.

Câu 16

An online retailer claims $p = 0.20$ of orders use a discount code. For an SRS of $n = 400$ orders, carry out a complete four-part check and find $P(\hat{p} > 0.25)$.

Lời giải chi tiết

(1) Center: $\mu_{\hat{p}} = 0.20$. **(2) Spread:** $\sigma_{\hat{p}} = \sqrt{\dfrac{0.20(0.80)}{400}} = \sqrt{0.0004} = 0.02$. **(3) Conditions:** Random (SRS given); 10% (the retailer surely processes far more than $4{,}000$ orders); Large Counts: $np = 80 \geq 10$, $n(1-p) = 320 \geq 10$ — all **met**, so $\hat{p}$ is approximately Normal. **(4) Probability:**

$$z = \frac{0.25 - 0.20}{0.02} = 2.5. \qquad P(\hat{p} > 0.25) = P(Z > 2.5) = 1 - 0.9938 = 0.0062.$$

An observed rate of 25% discount-code usage would be quite unusual (only about a 0.62% chance) if the true rate were really 20% — the kind of reasoning that becomes the basis of a significance test in later units.

Câu 17

Light bulb lifetimes have population standard deviation $\sigma = 100$ hours. A student wants $P(\bar{x} > 1000)$ for an SRS of $n = 25$ bulbs, but mistakenly uses $\sigma = 100$ directly in the z-score formula instead of the standard deviation of the sampling distribution. Explain the error and compute the correct value to use.

Lời giải chi tiết

The student has confused σ (the standard deviation of a *single* light bulb's lifetime) with $\sigma_{\bar{x}}$ (the standard deviation of the *sample mean* $\bar{x}$ of 25 bulbs) — these are different quantities, and $\sigma_{\bar{x}}$ is always smaller. The correct standard deviation to use is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{25}} = \frac{100}{5} = 20 \text{ hours},$$

which is five times smaller than $\sigma = 100$. Using σ in place of $\sigma_{\bar{x}}$ would make $\bar{x}$ appear far more variable than it actually is, distorting any resulting probability calculation — typically producing z-scores that are far too small in magnitude, and probabilities far too close to 0.5.

Câu 18

A population of grocery store transaction amounts is strongly right-skewed. For an SRS of $n = 45$ transactions, what can be said about the shape of the sampling distribution of $\bar{x}$?

(A) It is also strongly right-skewed, since the population is skewed

(B) It cannot be determined without knowing σ

(C) It is approximately Normal, because $n = 45 \geq 30$ satisfies the Central Limit Theorem, regardless of the population's skew

(D) It is exactly Normal only if $n \geq 100$

Lời giải chi tiết

By the Central Limit Theorem, once n is large enough (rule of thumb $n \geq 30$), the sampling distribution of $\bar{x}$ becomes approximately Normal *no matter what shape the population has*. Here $n = 45 \geq 30$, so the population's skew does not prevent $\bar{x}$ from being approximately Normal — only the population's own distribution stays skewed, never the sampling distribution of $\bar{x}$ itself. **(C)**.

Câu 19

A population of insurance claim amounts is strongly right-skewed. For an SRS of only $n = 5$ claims, can the sampling distribution of $\bar{x}$ be treated as approximately Normal?

(A) Yes, because $\bar{x}$ is always Normal regardless of n

(B) Yes, because $n \geq 2$ is enough for the CLT to apply

(C) No, because $n = 5$ is far too small for the CLT to overcome the population's strong skew, and the population itself is not Normal

(D) No, because σ is unknown

Lời giải chi tiết

Neither condition for Normality of $\bar{x}$ is met: the population is not Normal (it is strongly skewed), and $n = 5$ is far below the $n \geq 30$ threshold needed for the CLT to compensate for that skew. The sampling distribution of $\bar{x}$ here likely remains visibly skewed, so no z-score-based probability should be reported. **(C)**.

Câu 20

Insurance claims at a large firm are strongly right-skewed with population mean $\mu = 1200$ dollars and standard deviation $\sigma = 900$ dollars. For an SRS of $n = 100$ claims, find $P(\bar{x} > 1335)$.

Lời giải chi tiết

Conditions: Random (SRS given); 10% (a large firm processes far more than $1{,}000$ claims); population is skewed, not Normal, but $n = 100 \geq 30$, so by the CLT $\bar{x}$ is approximately Normal.

$$\sigma_{\bar{x}} = \frac{900}{\sqrt{100}} = 90. \qquad z = \frac{1335 - 1200}{90} = \frac{135}{90} = 1.5.$$

$$P(\bar{x} > 1335) = P(Z > 1.5) = 1 - 0.9332 = 0.0668.$$

Notice that the population's own standard deviation ($\$900$) is much larger than $\sigma_{\bar{x}}$ ($\$90$) — averaging 100 skewed claims together has already tamed the variability by a factor of 10, exactly as $\sigma/\sqrt{100} = \sigma/10$ predicts.

Câu 21

When the Random, 10%, and Large Counts conditions are all satisfied for a sample proportion $\hat{p}$, the sampling distribution of $\hat{p}$ is best described as:

(A) exactly binomial, and never treated as continuous

(B) approximately Normal – unimodal and symmetric, centered at p

(C) always right-skewed, regardless of the value of p

(D) uniform over all values between 0 and 1

Lời giải chi tiết

When Random, 10%, and Large Counts all hold, the (discrete) distribution of $\hat{p}$ is well approximated by a continuous, symmetric, bell-shaped Normal curve centered at $\mu_{\hat{p}} = p$. **(B)**.

Câu 22

Which statement about the Central Limit Theorem (CLT) is correct?

(A) The CLT guarantees that the population distribution itself becomes Normal as n increases

(B) The CLT applies only when the population distribution is already Normal

(C) The CLT states that the sampling distribution of $\bar{x}$ becomes approximately Normal as n increases, regardless of the population's shape

(D) The CLT guarantees $\sigma_{\bar{x}} = \sigma$ for any sample size n

Lời giải chi tiết

The CLT is a statement about the *sampling distribution of* $\bar{x}$, not about the population itself (the population's shape never changes). As n grows, $\bar{x}$'s sampling distribution becomes approximately Normal for essentially any population shape, and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ shrinks with n (it is never simply equal to σ for $n > 1$). **(C)**. Options (A) and (D) confuse the population with the sampling distribution of $\bar{x}$; option (B) reverses the entire point of the theorem, which is that Normality of the population is *not* required.

Câu 23

For a fixed population, to cut the standard deviation of the sampling distribution of $\bar{x}$ (or $\hat{p}$) exactly in half, the sample size n must be:

(A) doubled

(B) quadrupled (multiplied by 4)

(C) squared

(D) increased by 50%

Lời giải chi tiết

Since $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ (and similarly for $\sigma_{\hat{p}}$) depends on $1/\sqrt{n}$, halving the standard deviation requires $\sqrt{n_{\text{new}}} = 2\sqrt{n_{\text{old}}}$, i.e., $n_{\text{new}} = 4n_{\text{old}}$. **(B)**.

Câu 24

A population has standard deviation $\sigma = 12$. An SRS currently uses $n = 36$. By what factor must n increase to reduce $\sigma_{\bar{x}}$ to exactly one-third of its current value? Verify your answer numerically.

Lời giải chi tiết

Currently, $\sigma_{\bar{x}} = \dfrac{12}{\sqrt{36}} = \dfrac{12}{6} = 2$. To reduce this to one-third, we need the new standard deviation to equal $\dfrac{2}{3} \approx 0.667$. Since $\sigma_{\bar{x}} \propto 1/\sqrt{n}$, reducing $\sigma_{\bar{x}}$ by a factor of 3 requires increasing n by a factor of $3^2 = 9$. Check: $n_{\text{new}} = 9(36) = 324$, and $\sigma_{\bar{x}} = \dfrac{12}{\sqrt{324}} = \dfrac{12}{18} = 0.667$, which is exactly $\dfrac{1}{3}$ of the original value 2, as required. The sample size must increase **ninefold**, from $n = 36$ to $n = 324$ — confirming the general rule that reducing a sampling distribution's spread by a factor of k always costs a factor of k^2 in sample size. The same k^2 rule applies identically to $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, since both standard deviations depend on n only through $1/\sqrt{n}$.

Câu 25

For two independent SRSs used to estimate $\mu_1 - \mu_2$, the standard deviation of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is found by:

(A) adding σ_1 and σ_2 directly

(B) adding the two variances σ_1^2/n_1 and σ_2^2/n_2, then taking the square root

(C) subtracting the two variances

(D) multiplying σ_1 and σ_2

Lời giải chi tiết

Because the two samples are independent, their variances add: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Standard deviations are never added or subtracted directly — squaring, adding, then taking the square root at the end is required every time. **(B)**.

Câu 26

Vending machine snack weights are approximately Normal. Brand P has $\mu_1 = 45$ g, $\sigma_1 = 5$ g, with an SRS of $n_1 = 25$ bags tested. Brand Q has $\mu_2 = 42$ g, $\sigma_2 = 6$ g, with an independent SRS of $n_2 = 12$ bags tested. Find $P(\bar{x}_P - \bar{x}_Q > 1)$.

Lời giải chi tiết

Conditions: both samples random and independent (given); both populations stated to be approximately Normal, so the Normal condition holds regardless of the small sample sizes.

$$\mu_{\bar{x}_P - \bar{x}_Q} = 45 - 42 = 3.$$

$$\sigma_{\bar{x}_P - \bar{x}_Q} = \sqrt{\frac{5^2}{25} + \frac{6^2}{12}} = \sqrt{\frac{25}{25} + \frac{36}{12}} = \sqrt{1+3} = \sqrt{4} = 2.$$

$$z = \frac{1-3}{2} = -1.00. \qquad P(\bar{x}_P - \bar{x}_Q > 1) = P(Z > -1.00) = P(Z < 1.00) = 0.8413.$$

Câu 27

The standard deviation of a statistic's sampling distribution (such as $\sigma_{\bar{x}}$ or $\sigma_{\hat{p}}$) is also commonly known as the statistic's:

(A) margin of error
(B) standard error
(C) p-value
(D) confidence level

Lời giải chi tiết

The standard deviation of a sampling distribution is called the **standard error** of the statistic — terminology used throughout the confidence-interval and significance-testing units later in this course, where it appears directly inside every margin-of-error formula. **(B)**.

Câu 28

Consider the small population $\{5, 10, 15, 20\}$. List every possible SRS of size $n = 3$ (without replacement), compute $\bar{x}$ for each, and verify that the mean of the resulting sampling distribution equals the population mean.

Lời giải chi tiết

Population mean: $\mu = \dfrac{5+10+15+20}{4} = \dfrac{50}{4} = 12.5$. There are $\binom{4}{3} = 4$ equally likely samples: $\{5, 10, 15\} \to \bar{x} = 10$; $\{5, 10, 20\} \to \bar{x} \approx 11.67$; $\{5, 15, 20\} \to \bar{x} \approx 13.33$; $\{10, 15, 20\} \to \bar{x} = 15$. Mean of the sampling distribution: $\dfrac{10 + 11.67 + 13.33 + 15}{4} = \dfrac{50}{4} = 12.5$, exactly equal to the population mean $\mu = 12.5$ — confirming $\bar{x}$ is unbiased even with a sample size as large as $N - 1$ relative to the population.

Inference for Categorical Data: Proportions

Mục tiêu Unit: Khung suy diễn thống kê bốn bước (State – Plan – Do – Conclude); điều kiện Random – 10% – Large Counts cho suy diễn về tỉ lệ; khoảng tin cậy một tỉ lệ và cách xác định cỡ mẫu cần thiết cho một sai số biên cho trước; kiểm định giả thuyết một tỉ lệ (một phía và hai phía) và mối liên hệ giữa khoảng tin cậy với kiểm định hai phía; sai lầm loại I, sai lầm loại II, mức ý nghĩa α và năng lực kiểm định (power); khoảng tin cậy và kiểm định giả thuyết cho hiệu hai tỉ lệ $p_1 - p_2$ (phương pháp không gộp cho khoảng tin cậy, phương pháp gộp – pooled – cho kiểm định).

Many of the most consequential questions in medicine, business, politics, and public policy reduce to a single number: what proportion of some population has a certain characteristic, and does one group's proportion differ from another's? What fraction of a batch of vaccines is effective? What share of voters support a candidate? Does a redesigned checkout page convert visitors to purchases more often than the old one? A **categorical variable** sorts individuals into non-numeric groups (yes/no, defective/not defective, favor/oppose), and the **proportion** p of a population falling into a category of interest is the natural parameter for summarizing it. This unit develops the two core tools — confidence intervals and significance tests — used throughout statistics to answer such questions from sample data, together with the safeguards (Random, 10%, Large Counts) that keep the resulting conclusions trustworthy, and extends both tools to comparisons between two independent groups.

6.1 The Four-Step Framework for Statistical Inference

Every confidence interval and every significance test in this course can be organized around the same four steps. Using this framework consistently is not just good habit — on the AP exam, full credit on an inference free-response question requires each of these steps to appear explicitly.

State – Plan – Do – Conclude

STATE: Identify the population parameter of interest in context (a proportion p, or a difference in proportions $p_1 - p_2$). For a confidence interval, state the confidence level and what you intend to estimate. For a significance test, state the null and alternative hypotheses, H_0 and H_a, both in words and in symbols, and the significance level α.

PLAN: Name the specific inference procedure you will use (e.g., "one-sample z interval for a proportion," "two-sample z test for a difference in proportions"). Then verify *every* condition that procedure requires — **Random**, **10%**, **Large Counts** — using the actual numbers given in the problem. Simply writing "conditions are met" without checking each one earns no credit.

DO: If, and only if, the conditions are satisfied, carry out the mechanics: compute the sample statistic(s), the standard error (for a CI) or the standardized test statistic (for a test), and the resulting interval or P-value.

CONCLUDE: Answer the original question *in context*, using precise statistical language. For a test: "reject H_0" or "fail to reject H_0" — never "accept H_0" — justified by comparing the P-

value to α. For a confidence interval: "we are $C\%$ confident that the interval from ___ to ___ captures the true value of [parameter, described in context]."

VN **GIẢI THÍCH TIẾNG VIỆT**

Mọi bài toán suy diễn thống kê (khoảng tin cậy hoặc kiểm định giả thuyết) đều nên trình bày theo đúng bốn bước: **STATE** (Nêu) – xác định tham số tổng thể cần ước lượng hoặc nêu rõ H_0, H_a và mức ý nghĩa α; **PLAN** (Lập kế hoạch) – gọi tên đúng phương pháp suy diễn sẽ dùng và kiểm tra *đầy đủ* các điều kiện (Random, 10%, Large Counts) bằng số liệu cụ thể của đề bài, không chỉ viết chung chung "điều kiện đã thỏa"; **DO** (Thực hiện) – tính toán thống kê mẫu, sai số chuẩn (hoặc thống kê kiểm định) và khoảng tin cậy (hoặc P-value); **CONCLUDE** (Kết luận) – trả lời câu hỏi gốc trong *ngữ cảnh* đề bài, dùng đúng thuật ngữ: với kiểm định, chỉ được nói "bác bỏ H_0" hoặc "không bác bỏ H_0" (không bao giờ nói "chấp nhận H_0"); với khoảng tin cậy, phải nói "chúng ta tin tưởng $C\%$ rằng khoảng này chứa giá trị thật của...". Đề thi AP luôn chấm điểm theo từng bước riêng biệt, nên bỏ sót một bước (dù tính toán đúng) vẫn bị trừ điểm.

6.2 Conditions for Inference about a Proportion

6.2.1 The Sampling Distribution of a Sample Proportion

Before any inference procedure can be justified, it helps to know how the sample proportion $\hat{p} = x/n$ (where x is the count of successes in the sample) behaves across repeated random samples of size n drawn from a population with true proportion p. Three facts describe this **sampling distribution**:

- **Center:** $\hat{p}$ is an **unbiased** estimator of p — averaged across all possible samples of size n, $\mu_{\hat{p}} = p$. Some samples will overestimate p and some will underestimate it, but there is no systematic tendency in either direction.
- **Spread:** the standard deviation of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, which **decreases** as n increases — larger samples give more precise estimates of p. Unlike inference for a population mean, this standard deviation depends only on p and n; there is no separate, unknown population standard deviation to estimate from the data, so no t-distribution is required here. z-procedures apply directly, with $\hat{p}$ (for a CI) or p_0 (for a test) substituted in place of the unknown p.
- **Shape:** by the Central Limit Theorem applied to a proportion (equivalently, the Normal approximation to the Binomial distribution), the sampling distribution of $\hat{p}$ is approximately Normal *provided the sample is large enough* — this is exactly the role played by the Large Counts condition below.

Together, these three facts justify every z-based procedure in this unit: once Random, 10%, and Large Counts all hold, $\hat{p}$ behaves like a draw from an approximately Normal distribution centered at p with standard deviation $\sqrt{p(1-p)/n}$, and the familiar z^* and z-statistic machinery follows directly.

6.2.2 The Random, 10% and Large Counts Conditions

Random: the data must come from a random sample or, in an experiment, from random assignment to treatment groups. Without this, no probability-based method of inference can legitimately extend conclusions beyond the individuals actually observed.

10% condition: when sampling *without replacement* from a finite population of size N, require $n \leq 0.10N$. Removing individuals from a finite population without replacement technically makes successive observations dependent (the composition of what remains changes slightly with each draw); the 10% condition keeps this dependence small enough that the standard variance/standard-error formulas — which assume independence — remain approximately valid.

Large Counts condition (also called the Normal or large-sample condition): the sampling distribution of $\hat{p}$ is only approximately Normal when the expected number of successes and the expected number of failures are both at least 10. Critically, *which* proportion you plug in depends on the procedure: for a **confidence interval** (where the true p is unknown), check $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ using the *sample* proportion $\hat{p}$; for a **significance test** (where H_0 specifies a hypothesized value), check $np_0 \geq 10$ and $n(1-p_0) \geq 10$ using the *hypothesized* proportion p_0, since a test temporarily assumes H_0 is true.

VN **GIẢI THÍCH TIẾNG VIỆT**

Ba điều kiện bắt buộc trước khi dùng phương pháp z cho tỉ lệ: **Random** – dữ liệu phải đến từ mẫu ngẫu nhiên (hoặc phân bổ ngẫu nhiên trong thí nghiệm); **10%** – nếu lấy mẫu *không hoàn lại* từ một tổng thể hữu hạn kích thước N, cỡ mẫu n phải không vượt quá 10% của N, để các lần chọn vẫn xấp xỉ độc lập với nhau; **Large Counts** – phân phối lấy mẫu của $\hat{p}$ chỉ xấp xỉ chuẩn khi số "thành công" và số "thất bại" kỳ vọng đều ≥ 10. Điểm dễ nhầm nhất: với **khoảng tin cậy** thì kiểm tra bằng $\hat{p}$ (giá trị mẫu, vì p thật chưa biết); còn với **kiểm định giả thuyết** thì kiểm tra bằng p_0 (giá trị giả thuyết trong H_0, vì phép kiểm định tạm giả sử H_0 đúng để tính toán). Nhầm lẫn hai công thức này là lỗi rất phổ biến khi kiểm tra điều kiện.

6.2.3 Checking Conditions in Practice

Ví dụ mẫu · Conditions all satisfied

A quality-control manager at an electronics distributor wants to construct a confidence interval for the true proportion of defective units in a shipment of $N = 5{,}000$ items. She selects a simple random sample of $n = 150$ items and finds 18 defective. Check whether the conditions for a one-proportion z interval are satisfied.

Random: an SRS was selected — satisfied.
10%: $0.10N = 0.10(5{,}000) = 500$; the sample size $n = 150 \leq 500$ — satisfied.
Large Counts: $\hat{p} = 18/150 = 0.12$; $n\hat{p} = 150(0.12) = 18 \geq 10$ and $n(1-\hat{p}) = 150(0.88) = 132 \geq 10$ — satisfied.
All three conditions hold, so a one-proportion z interval is appropriate here.

Ví dụ mẫu · The 10% condition fails

A regional hobby association has $N = 300$ members. An officer surveys $n = 40$ members, selected at random and *without replacement*, about a proposed rule change; 18 support it. Check the conditions for a one-proportion z procedure.

Random: an SRS is stated — satisfied.

10%: $0.10N = 0.10(300) = 30$; but $n = 40 > 30$ — **this condition fails.** Sampling more than 10% of a small, finite population without replacement means successive selections are no longer close enough to independent for the standard standard-error formula to be trustworthy.
Large Counts: $\hat{p} = 18/40 = 0.45$; $n\hat{p} = 18 \geq 10$ and $n(1-\hat{p}) = 22 \geq 10$ — this condition alone would be satisfied.
Because the 10% condition fails, a standard one-proportion z procedure should not be used here without a finite-population correction, a larger population, or a smaller sample.

6.3 Confidence Intervals for a Proportion

General form: statistic $\pm$ (critical value)$\times$(standard deviation of statistic). For one proportion:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

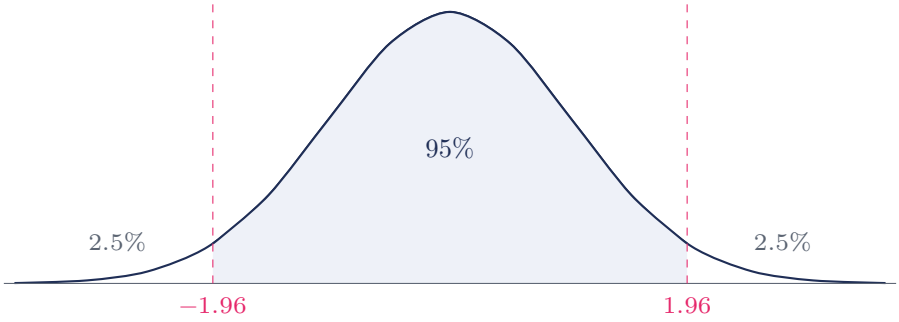
Conditions: **Random** (SRS or random assignment); **10%** condition ($n \leq 0.10N$ if sampling without replacement); **Large Counts**: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$. Common critical values: $90\% \to z^* = 1.645$; $95\% \to 1.960$; $99\% \to 2.576$.

Ví dụ mẫu · Worked example

A random sample of $n = 200$ likely voters finds 118 favor a ballot measure. Construct and interpret a 95% confidence interval for the true proportion p who favor it.
$\hat{p} = \dfrac{118}{200} = 0.59$. **Conditions:** random sample given; $n = 200$ is surely $\leq 10\%$ of all voters; Large Counts: $n\hat{p} = 118 \geq 10$, $n(1-\hat{p}) = 82 \geq 10$ — **conditions met.**

$$SE = \sqrt{\frac{0.59(0.41)}{200}} = \sqrt{0.0012095} \approx 0.0348, \qquad ME = 1.96(0.0348) \approx 0.068.$$

95% CI: $0.59 \pm 0.068 = (0.522,\ 0.658)$. *Interpretation:* we are 95% confident the true proportion of all likely voters favoring the measure is between 52.2% and 65.8%.



The standard Normal curve with the middle 95% of area shaded ($z^* = \pm 1.96$). A 95% confidence interval is built by extending a margin of error of $z^* \cdot SE$ on either side of $\hat{p}$, so that the method captures the true p in about 95% of all possible random samples.

6.3.1 Choosing a Sample Size for a Desired Margin of Error

Before collecting data, researchers often need to know how large a sample must be to guarantee a margin of error no larger than some target value ME. Starting from $ME = z^*\sqrt{p^*(1-p^*)/n}$ and

solving for n:

$$n = \left(\frac{z^*}{ME}\right)^2 p^*(1-p^*),$$

where p^* is a planning value for p — either a reasonable estimate from prior data, or, if no such estimate exists, the **conservative** choice $p^* = 0.5$. Because $p^*(1-p^*)$ is maximized at $p^* = 0.5$, using 0.5 guarantees the actual margin of error will be at most ME no matter what the true p turns out to be — but it also produces the *largest possible* required sample size. Always round n **up** to the next whole number, since rounding down would allow the margin of error to exceed the target.

Ví dụ mẫu · Determining sample size

A marketing team wants to estimate, with 95% confidence and a margin of error of at most 0.03 (3 percentage points), the proportion of customers who would recommend a new product.

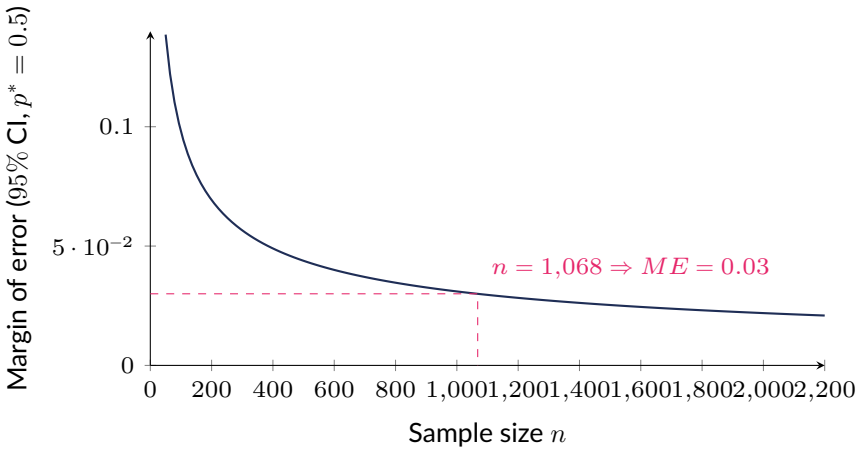
(a) No prior estimate available. Use the conservative $p^* = 0.5$:

$$n = \left(\frac{1.96}{0.03}\right)^2 (0.5)(0.5) = \frac{(1.96)^2(0.25)}{(0.03)^2} = \frac{0.9604}{0.0009} \approx 1067.1 \;\Rightarrow\; n = 1{,}068.$$

(b) A pilot survey suggests about 20% of customers would recommend it. Use $p^* = 0.20$:

$$n = \frac{(1.96)^2(0.20)(0.80)}{(0.03)^2} = \frac{0.614656}{0.0009} \approx 682.9 \;\Rightarrow\; n = 683.$$

Using the realistic prior estimate ($p^* = 0.20$) nearly halves the required sample size compared to the conservative default (683 vs. $1{,}068$), because $0.5(1-0.5) = 0.25$ is larger than $0.2(0.8) = 0.16$. The conservative guess is always "safe" but never cheaper.



Margin of error for a 95% one-proportion z interval as a function of n, using the conservative planning value $p^* = 0.5$. ME decreases as n increases, but at a diminishing rate — quadrupling n only halves ME, since $ME \propto 1/\sqrt{n}$.

6.3.2 Confidence Level versus a Confidence Interval

Confidence level (e.g., "95%") describes the long-run success rate of the *method* used to build the interval: if we repeated the random sampling process many times and constructed a confidence interval from each sample using the same procedure, approximately $C\%$ of those intervals would capture the true parameter value.

A specific confidence interval already calculated from one sample (e.g., "$(0.522, 0.658)$") either does or does not contain the true parameter — once the data are collected there is nothing left to be random. It is therefore **incorrect** to say "there is a 95% probability that p is in this interval." The correct statement is: "we are $C\%$ **confident** that this interval captures the true value of [parameter, in context]" — the confidence describes trust in the method, applied to this one outcome of it.

VN **GIẢI THÍCH TIẾNG VIỆT**

Cần phân biệt hai cách diễn giải: **mức tin cậy** (95%) mô tả *tỉ lệ thành công về lâu dài của phương pháp* – nếu lặp lại việc lấy mẫu và xây dựng khoảng tin cậy nhiều lần theo đúng quy trình, khoảng 95% số khoảng đó sẽ chứa giá trị thật p. Ngược lại, **một khoảng tin cậy cụ thể** đã tính được (ví dụ $(0.522, 0.658)$) thì hoặc có chứa p hoặc không – không còn gì ngẫu nhiên nữa một khi đã có số liệu, nên KHÔNG được nói "có 95% xác suất p nằm trong khoảng này". Câu đúng: "chúng ta tin tưởng 95% rằng khoảng này chứa giá trị thật của p" – niềm tin đặt vào *phương pháp*, không phải vào một xác suất áp dụng cho khoảng cụ thể đó.

A note on small samples. The standard $\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$ interval tends to perform poorly — its true capture rate falls noticeably below the stated confidence level — when n is small or when $\hat{p}$ is close to 0 or 1, even if the Large Counts condition is technically satisfied. A well-known refinement, the **plus-four (Agresti–Coull) interval**, adds two artificial successes and two artificial failures before computing: use $\tilde{p} = \dfrac{x+2}{n+4}$ in place of $\hat{p}$, and $n+4$ in place of n, in the same formula. This adjustment is not required for AP-level work (where the Large Counts condition is treated as sufficient), but it explains why some statistical software reports slightly different interval bounds than the by-hand z-interval taught in this course.

6.4 Significance Tests for a Proportion

Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$, using the *hypothesized* p_0 in the standard deviation (not $\hat{p}$).

Ví dụ mẫu · Worked example

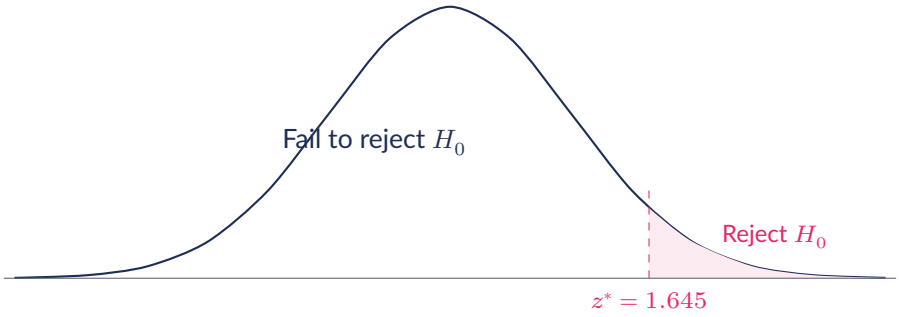
Test, using the voter data above ($\hat{p} = 0.59$, $n = 200$), whether a *majority* favor the measure: $H_0 : p = 0.5$ vs. $H_a : p > 0.5$, at $\alpha = 0.05$.

$$z = \frac{0.59 - 0.50}{\sqrt{0.5(0.5)/200}} = \frac{0.09}{\sqrt{0.00125}} = \frac{0.09}{0.03536} \approx 2.55.$$

P-value $= P(Z > 2.55) \approx 0.0055$. Since $0.0055 < 0.05$, **reject** H_0: there is convincing evidence that a majority of likely voters favor the measure.

GIẢI THÍCH TIẾNG VIỆT VN

Với kiểm định một tỉ lệ, mẫu số dùng p_0 (tỉ lệ giả thuyết) chứ **không** dùng $\hat{p}$ — khác với khoảng tin cậy (dùng $\hat{p}$ vì p chưa biết). "Bác bỏ H_0" nghĩa là có bằng chứng thuyết phục ủng hộ H_a; "không bác bỏ H_0" **không có nghĩa là** H_0 đúng, chỉ là chưa đủ bằng chứng.



Rejection region for a one-sided upper-tail test at $\alpha = 0.05$: reject H_0 if the standardized test statistic falls at or beyond $z^* = 1.645$ (the shaded area equals α).

6.4.1 A One-Sided Test in the Other Direction

Ví dụ mẫu · A "less than" alternative

A shipping company claims that at least 90% of its packages are delivered on time ($p_0 = 0.90$). A regulator suspects the true on-time rate is actually lower and randomly samples $n = 120$ recent packages, finding that 100 were delivered on time. Test at $\alpha = 0.05$.

State: let $p =$ the true proportion of all packages delivered on time. $H_0 : p = 0.90$; $H_a : p < 0.90$; $\alpha = 0.05$.

Plan: one-sample z test for a proportion. *Random* — SRS given. *10%* — 120 packages is far less than 10% of all packages the company ships. *Large Counts (test, use p_0):* $np_0 = 120(0.90) = 108 \geq 10$; $n(1-p_0) = 120(0.10) = 12 \geq 10$ — conditions met.

Do: $\hat{p} = \dfrac{100}{120} \approx 0.8333$.

$$SE = \sqrt{\frac{0.90(0.10)}{120}} = \sqrt{0.00075} \approx 0.02739, \qquad z = \frac{0.8333 - 0.90}{0.02739} \approx -2.43.$$

P-value $= P(Z < -2.43) = 1 - \Phi(2.43)$. Interpolating in the standard Normal table between $2.40 \to 0.9918$ and $2.50 \to 0.9938$: $\Phi(2.43) \approx 0.9918 + 0.3(0.0020) = 0.9924$, so P-value $\approx 1 - 0.9924 = 0.0076$.

Conclude: since $0.0076 < 0.05$, **reject** H_0. There is convincing evidence that the true on-time delivery rate for this company is below 90%.

6.4.2 A Two-Sided Test

Ví dụ mẫu · A two-sided test

A snack company states that 8% of its bags contain a "prize ticket" ($p_0 = 0.08$). A consumer group tests whether the true proportion differs from 8%, sampling $n = 400$ bags at random and finding 25 with tickets. Test at $\alpha = 0.05$.

State: let $p =$ the true proportion of all bags containing a prize ticket. $H_0 : p = 0.08$; $H_a : p \neq 0.08$; $\alpha = 0.05$.

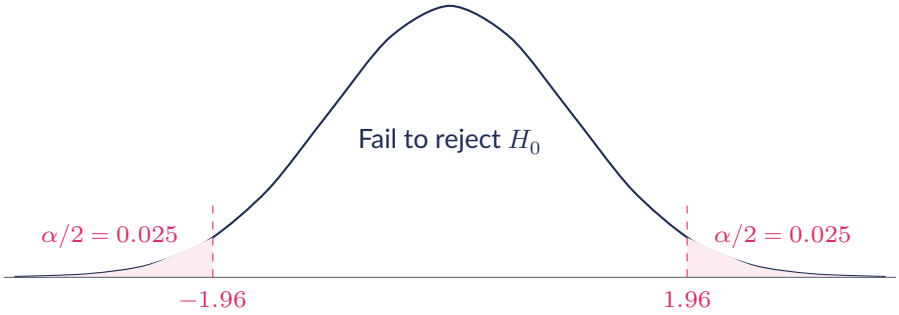
Plan: one-sample z test for a proportion. *Random* — SRS given. *10%* — 400 bags is a tiny fraction of total production. *Large Counts (use p_0):* $np_0 = 400(0.08) = 32 \geq 10$; $n(1-p_0) = 400(0.92) = 368 \geq 10$ — conditions met.

Do: $\hat{p} = \dfrac{25}{400} = 0.0625$.

$$SE = \sqrt{\frac{0.08(0.92)}{400}} = \sqrt{0.000184} \approx 0.01357, \qquad z = \frac{0.0625 - 0.08}{0.01357} \approx -1.29.$$

Two-sided P-value $= 2P(Z < -1.29) = 2\,[1 - \Phi(1.29)]$. Interpolating between $1.28 \to 0.8997$ and $1.30 \to 0.9032$: $\Phi(1.29) \approx 0.9015$, so P-value $\approx 2(1 - 0.9015) = 2(0.0985) = 0.197$.

Conclude: since $0.197 > 0.05$, **fail to reject** H_0. There is not convincing evidence that the true proportion of bags with a prize ticket differs from 8%.



Rejection region for a two-sided test at $\alpha = 0.05$: reject H_0 if $z \leq -1.96$ or $z \geq 1.96$ — the two shaded tails together total area α.

6.4.3 Connecting the Confidence Interval and the Two-Sided Test

A useful check: a two-sided test of $H_0 : p = p_0$ at significance level α will **reject** H_0 exactly when the corresponding $(1-\alpha)$ confidence interval for p does **not** contain p_0 — the two procedures are built from the same information and must agree.

Ví dụ mẫu · Do the interval and the test agree?

Using the prize-ticket data above ($\hat{p} = 0.0625$, $n = 400$), construct a 95% confidence interval for p and check whether it is consistent with the two-sided test's conclusion.

$$SE = \sqrt{\frac{0.0625(0.9375)}{400}} = \sqrt{0.0001465} \approx 0.01210, \qquad ME = 1.96(0.01210) \approx 0.0237.$$

95% CI: $0.0625 \pm 0.0237 = (0.0388,\ 0.0862)$.

The hypothesized value $p_0 = 0.08$ **does** fall inside $(0.0388, 0.0862)$. This is exactly consistent with the earlier two-sided test, which **failed to reject** $H_0 : p = 0.08$ at $\alpha = 0.05$ — whenever p_0 lies inside the 95% CI, the corresponding two-sided test at $\alpha = 0.05$ will fail to reject. Notice the CI's standard error used $\hat{p} = 0.0625$ (the sample value), while the test's standard error used $p_0 = 0.08$ (the hypothesized value) — the two SE formulas are related but not identical, which is why the numeric boundary values differ slightly from a pure algebraic inversion of the test.

6.4.4 Interpreting P-values and Statistical Significance

A P**-value** is the probability, computed *assuming* H_0 *is true*, of obtaining a sample statistic at least as extreme (in the direction of H_a) as the one actually observed. A small P-value means the observed result would be surprising if H_0 were true — evidence against H_0. Comparing the P-value to a pre-chosen significance level α gives a firm decision rule: if $P < \alpha$, the result is **statistically significant** at level α and we reject H_0; if $P \geq \alpha$, we fail to reject H_0.

Statistical significance is not the same as practical significance. Because $SE = \sqrt{p(1-p)/n}$ shrinks as n grows, a sufficiently large sample can make even a tiny, practically unimportant difference from p_0 produce a very small P-value. Conversely, a real and practically meaningful difference can fail to reach statistical significance if the sample is small (low power). Always report and interpret the size of the estimated effect itself (the point estimate, or better, a confidence interval), not merely whether $P < \alpha$.

Ví dụ mẫu · Statistical versus practical significance

An online retailer runs an enormous study with $n = 1{,}000{,}000$ users, testing whether a new checkout-button color changes the historical conversion rate $p_0 = 0.500$. It observes $\hat{p} = 0.505$ and tests $H_0 : p = 0.500$ vs. $H_a : p \neq 0.500$.

$$SE = \sqrt{\frac{0.5(0.5)}{1{,}000{,}000}} = 0.0005, \qquad z = \frac{0.505 - 0.500}{0.0005} = 10.$$

A z-statistic this large corresponds to an astronomically small two-sided P-value ($P \approx 0$), so the result is overwhelmingly **statistically significant** — reject H_0 with essentially total confidence. Yet the estimated effect itself — a shift from 50.0% to 50.5% conversion — may be far too small to matter for a business decision. With n this large, even a trivial, practically meaningless difference from p_0 becomes statistically detectable, which is exactly why the size of the effect (not just significance) must always be examined.

6.5 Type I and Type II Errors, Significance Level, and Power

Type I error: rejecting H_0 when H_0 is actually true (a "false positive"). For a test conducted at significance level α, $P(\text{Type I error}) = \alpha$ whenever H_0 is true.

Type II error: failing to reject H_0 when H_0 is actually false (a "false negative"); its probability is denoted β.

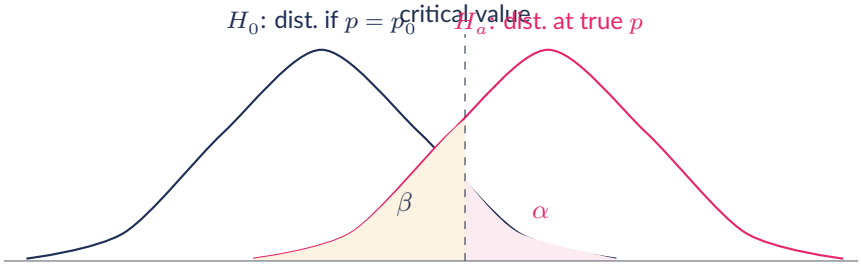
Power of a test $= 1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ is false, for a specific true parameter value})$ — the probability the test correctly detects a real effect of that size.

Power **increases** when: (i) the sample size n increases (smaller SE, sharper separation between the H_0 and H_a sampling distributions); (ii) the true parameter value is farther from p_0 (a larger effect size is easier to detect); (iii) the significance level α is increased (a larger rejection region makes H_0 easier to reject, but at the cost of a higher Type I error rate). For fixed n and effect size, α and β trade off: shrinking α enlarges β (lowers power), and vice versa.

VN **GIẢI THÍCH TIẾNG VIỆT**

Sai lầm loại I (Type I error): bác bỏ H_0 trong khi H_0 thực ra đúng ("dương tính giả"); xác suất mắc sai lầm này chính là α. **Sai lầm loại II** (Type II error): không bác bỏ H_0 trong khi H_0 thực ra sai ("âm tính giả"); xác suất là β. **Năng lực kiểm định (power)** $= 1 - \beta$ là xác suất phát hiện

đúng một hiệu ứng có thật. Power tăng khi: (1) cỡ mẫu n tăng (sai số chuẩn nhỏ hơn, hai phân phối lấy mẫu tách biệt rõ hơn); (2) hiệu ứng thật (khoảng cách giữa p thật và p_0) lớn hơn; (3) mức ý nghĩa α tăng (vùng bác bỏ rộng hơn, nhưng đổi lại tăng nguy cơ sai lầm loại I). Với n và hiệu ứng cố định, α và β luôn đánh đổi lẫn nhau: giảm α sẽ làm tăng β (giảm power).



Conceptual power diagram: the navy curve is the sampling distribution of $\hat{p}$ assuming H_0 is true; the pink curve is the sampling distribution at the true (alternative) value of p. The dashed line marks the test's critical value. α (pink-shaded) is the area of the H_0 curve beyond the critical value; β (amber-shaded) is the area of the H_a curve on the "fail to reject" side. Power $= 1 - \beta$ is the remaining area of the H_a curve beyond the critical value.

Ví dụ mẫu · Type I and Type II errors in context

A factory targets a defect rate of at most 2%. Quality control tests $H_0 : p = 0.02$ (the defect rate meets target) vs. $H_a : p > 0.02$ (the rate has risen above target, triggering a line inspection).

Type I error: concluding the defect rate exceeds 2% (halting the line for inspection) when the true rate is actually still 2%. *Consequence:* an unnecessary, costly production stoppage even though nothing was actually wrong.

Type II error: failing to detect that the defect rate has truly risen above 2% (allowing production to continue unchanged) when it has, in fact, increased. *Consequence:* the factory keeps shipping a higher-than-intended proportion of defective units to customers, risking safety, warranty costs, and reputational damage before the problem is eventually caught.

Ví dụ mẫu · Power increases with sample size

A pollster tests $H_0 : p = 0.5$ vs. $H_a : p > 0.5$ at $\alpha = 0.05$ ($z^* = 1.645$), asking whether more than half of voters support a candidate. Suppose, unknown to the pollster, the true support is $p = 0.65$. Compare the power of this test using (a) $n = 25$ and (b) $n = 100$.

(a) $n = 25$. Critical value of $\hat{p}$: $SE_0 = \sqrt{0.5(0.5)/25} = 0.10$, so $\hat{p}^* = 0.5 + 1.645(0.10) = 0.6645$. Under the true $p = 0.65$: $SE_a = \sqrt{0.65(0.35)/25} = \sqrt{0.0091} \approx 0.0954$.

$$z_{\text{power}} = \frac{0.6645 - 0.65}{0.0954} \approx 0.15, \qquad \text{Power} = P(Z \geq 0.15) = 1 - \Phi(0.15).$$

Interpolating between $0.00 \to 0.5000$ and $0.25 \to 0.5987$: $\Phi(0.15) \approx 0.5000 + 0.6(0.0987) = 0.559$. Power $\approx 1 - 0.559 = 0.44$ (about 44%).

(b) $n = 100$. $SE_0 = \sqrt{0.5(0.5)/100} = 0.05$, so $\hat{p}^* = 0.5 + 1.645(0.05) = 0.58225$. Under $p = 0.65$: $SE_a = \sqrt{0.65(0.35)/100} \approx 0.04770$.

$$z_{\text{power}} = \frac{0.58225 - 0.65}{0.04770} \approx -1.42, \qquad \text{Power} = P(Z \geq -1.42) = \Phi(1.42).$$

Interpolating between $1.40 \to 0.9192$ and $1.50 \to 0.9332$: $\Phi(1.42) \approx 0.9192 + 0.2(0.0140) = 0.922$. Power ≈ 0.92 (about 92%).

Quadrupling the sample size ($25 \to 100$) raises the power to detect this same true effect ($p = 0.65$ vs. hypothesized 0.5) from roughly 44% to roughly 92% — a dramatic illustration of how larger samples make a test far more likely to detect a real effect of a given size.

Ví dụ mẫu · Type I and Type II error in a two-sample context

A researcher compares pass rates between two teaching methods using $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$. Describe, in context, what a Type I error and a Type II error would mean here.

A **Type I error** means concluding the two teaching methods have different true pass rates when, in fact, their true pass rates are equal. *Consequence:* a school might invest significant time and money adopting the "better" method campus-wide based on a difference that does not actually exist, while an equally effective (and perhaps cheaper or simpler) alternative is abandoned for no real reason. A **Type II error** means concluding there is no convincing evidence of a difference when, in fact, one method's true pass rate genuinely is higher. *Consequence:* the school keeps using an inferior method and students continue to be under-served, simply because the study did not have enough evidence (often due to too small a sample) to detect a real, meaningful difference.

6.6 Comparing Two Proportions: A Confidence Interval for a Difference

For independent random samples (or independent groups) from two populations, the confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

This is the **unpooled** standard error: each sample uses its own proportion, because a confidence interval does not assume $p_1 = p_2$ — it is meant to estimate whatever the true difference actually is. **Conditions:** Random for each sample/group; 10% for each sample if sampled without replacement; the two samples must be **independent** of one another; Large Counts for *each* sample: $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, $n_2(1-\hat{p}_2) \geq 10$.

Independent samples, not matched pairs. The two-sample procedures in this section (and the pooled test in the next) require the two samples or groups to be genuinely **independent** — no individual should appear in, or be linked to, both groups. A common design that violates this is a **matched-pairs** or **before/after** study, where the same individuals (or naturally paired individuals, such as siblings or matched patients) are measured under both conditions. Applying an independent two-sample z procedure to matched data is a genuine error: it ignores the correlation between paired observations, typically producing a standard error that is too large and making real effects harder to detect. Recognizing whether a study design yields independent groups or paired/matched observations is itself part of the **Plan** step's condition-checking.

Ví dụ mẫu · Confidence interval for a difference in proportions

Independent random samples of customers are taken from two branches of a retail chain. Branch 1: $n_1 = 120$ surveyed, 96 report being satisfied. Branch 2: $n_2 = 150$ surveyed, 108 report being satisfied. Construct and interpret a 95% confidence interval for $p_1 - p_2$.

State: let $p_1, p_2 =$ the true proportions of satisfied customers at Branch 1 and Branch 2. Estimate $p_1 - p_2$ with 95% confidence.

Plan: two-sample z interval for a difference in proportions (unpooled). *Random* — independent SRSs given. *10%* — both samples are small relative to each branch's customer base. *Independent samples* — different branches, unrelated customers. *Large Counts:* $n_1\hat{p}_1 = 96 \geq 10$, $n_1(1-\hat{p}_1) = 24 \geq 10$; $n_2\hat{p}_2 = 108 \geq 10$, $n_2(1-\hat{p}_2) = 42 \geq 10$ — conditions met.

Do: $\hat{p}_1 = \dfrac{96}{120} = 0.80$; $\hat{p}_2 = \dfrac{108}{150} = 0.72$; point estimate $\hat{p}_1 - \hat{p}_2 = 0.08$.

$$SE = \sqrt{\frac{0.80(0.20)}{120} + \frac{0.72(0.28)}{150}} = \sqrt{0.001333 + 0.001344} = \sqrt{0.002677} \approx 0.0517.$$

$$ME = 1.96(0.0517) \approx 0.1014. \qquad 95\%\ \text{CI}: \ 0.08 \pm 0.1014 = (-0.021,\ 0.181).$$

Conclude: we are 95% confident that the true difference in satisfaction proportions (Branch 1 minus Branch 2) is between -0.021 and 0.181. Because this interval contains 0, it is plausible that there is no real difference between the two branches' satisfaction rates, even though the interval leans positive.

Ví dụ mẫu · A two-sample CI that does not contain zero

Two tutoring programs report pass rates on a certification exam. Program A: $n_1 = 200$ students, 170 pass. Program B: $n_2 = 180$ students, 126 pass. Construct a 95% confidence interval for $p_A - p_B$ and interpret it.

$\hat{p}_A = 170/200 = 0.85$; $\hat{p}_B = 126/180 = 0.70$. Conditions (Random, 10%, independent groups, Large Counts: $170, 30, 126, 54$ all ≥ 10) are satisfied.

$$SE = \sqrt{\frac{0.85(0.15)}{200} + \frac{0.70(0.30)}{180}} = \sqrt{0.0006375 + 0.0011667} = \sqrt{0.0018042} \approx 0.0425.$$

$$ME = 1.96(0.0425) \approx 0.0833. \qquad 95\%\ \text{CI}: \ 0.15 \pm 0.0833 = (0.067,\ 0.233).$$

Unlike the branch-satisfaction example above, this interval is **entirely positive** — it does not contain 0. We are 95% confident the true difference $p_A - p_B$ is between 0.067 and 0.233, which is convincing evidence that Program A's true pass rate genuinely is higher than Program B's, not merely a difference attributable to sampling variability.

6.7 Comparing Two Proportions: A Significance Test Using Pooled Proportions

For two proportions, use the pooled estimate $\hat{p}_c = \dfrac{x_1 + x_2}{n_1 + n_2}$ under $H_0 : p_1 = p_2$:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Pooling is appropriate here *because* H_0 assumes the two population proportions are equal, so the best single estimate of that common proportion combines both samples. **Large Counts for the test:** check $n_1\hat{p}_c \geq 10$, $n_1(1-\hat{p}_c) \geq 10$, $n_2\hat{p}_c \geq 10$, $n_2(1-\hat{p}_c) \geq 10$, using the *pooled* proportion in place of each sample's own $\hat{p}$.

Ví dụ mẫu · Two-sample test using the pooled proportion

An agricultural researcher compares seed germination rates for two fertilizers. New fertilizer: $n_1 = 160$ seeds treated, 128 germinate. Standard fertilizer: $n_2 = 180$ seeds treated, 126 germinate. Test whether the new fertilizer's true germination rate is higher, at $\alpha = 0.05$.

State: let p_1, p_2 = the true germination proportions under the new and standard fertilizers. $H_0 : p_1 = p_2$ (equivalently $p_1 - p_2 = 0$); $H_a : p_1 > p_2$; $\alpha = 0.05$.

Plan: two-sample z test for a difference in proportions (pooled). *Random* — random assignment of seeds to treatment assumed. *10%* — samples are small relative to any realistic population of seeds. *Independent groups* — the two treatment groups do not overlap. *Large Counts (pooled):* $\hat{p}_c = \dfrac{128+126}{160+180} = \dfrac{254}{340} \approx 0.7471$; $n_1\hat{p}_c \approx 119.5 \geq 10$, $n_1(1-\hat{p}_c) \approx 40.5 \geq 10$, $n_2\hat{p}_c \approx 134.5 \geq 10$, $n_2(1-\hat{p}_c) \approx 45.5 \geq 10$ — conditions met.

Do: $\hat{p}_1 = \dfrac{128}{160} = 0.80$; $\hat{p}_2 = \dfrac{126}{180} = 0.70$.

$$SE = \sqrt{0.7471(0.2529)\left(\frac{1}{160}+\frac{1}{180}\right)} = \sqrt{0.18896(0.011806)} = \sqrt{0.002231} \approx 0.04723.$$

$$z = \frac{0.80 - 0.70}{0.04723} \approx 2.12.$$

P-value $= P(Z > 2.12) = 1 - \Phi(2.12)$. Interpolating between $2.10 \to 0.9821$ and $2.20 \to 0.9861$: $\Phi(2.12) \approx 0.9821 + 0.2(0.0040) = 0.9829$, so P-value $\approx 1 - 0.9829 = 0.017$.

Conclude: since $0.017 < 0.05$, **reject** H_0. There is convincing evidence that the true germination rate under the new fertilizer is higher than under the standard fertilizer.

6.8 Formula Summary, Technology, and Choosing a Procedure

6.8.1 Formula Summary

Procedure	Formula	Key conditions
One-prop. CI	$\hat{p} \pm z^*\sqrt{\hat{p}(1-\hat{p})/n}$	Random, 10%, $n\hat{p} \geq 10$, $n(1-\hat{p}) \geq 10$
Sample size	$n = (z^*/ME)^2 p^*(1-p^*)$, round up	use $p^* = 0.5$ if no prior estimate
One-prop. test	$z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$	Random, 10%, $np_0 \geq 10$, $n(1-p_0) \geq 10$
Two-prop. CI	$(\hat{p}_1 - \hat{p}_2) \pm z^*\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	Random, 10%, independent samples, Large Counts per sample (unpooled)
Two-prop. test	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, $\hat{p}_c = \dfrac{x_1+x_2}{n_1+n_2}$	Random, 10%, independent samples, Large Counts using pooled $\hat{p}_c$

Critical values: $90\% \to z^* = 1.645$; $95\% \to 1.960$; $98\% \to 2.326$; $99\% \to 2.576$.

Using technology. Graphing calculators and statistical software implement each procedure above directly: on a TI-84, the `1-PropZInt` and `1-PropZTest` menus compute the one-proportion CI and test from x, n, and (for the test) p_0; `2-PropZInt` and `2-PropZTest` do the same for two samples, automatically using the unpooled SE for the interval and the pooled SE for the test. On the AP exam, using technology to carry out the mechanical **Do** step is fully acceptable, but the **Plan** step — naming the procedure and explicitly checking every condition — must still be written out; a calculator cannot verify Random, 10%, or Large Counts for you.

Ví dụ mẫu · Reading computer output

Statistical software reports the following for a two-proportion z-test comparing click-through rates on two website banner designs:

Sample 1	$n_1 = 500,\ x_1 = 42\ \ (\hat{p}_1 = 0.084)$
Sample 2	$n_2 = 500,\ x_2 = 25\ \ (\hat{p}_2 = 0.050)$
Pooled proportion	$\hat{p}_c = 0.067$
Test statistic	$z = 2.15$
P-value	0.032

State the hypotheses being tested and give a conclusion at $\alpha = 0.05$.

This output corresponds to $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$ (a two-sided test — software reports a two-tailed P-value by default unless a direction is specified). Checking the arithmetic: $\hat{p}_c = (42 + 25)/(500 + 500) = 67/1000 = 0.067$, and $SE = \sqrt{0.067(0.933)(1/500 + 1/500)} \approx 0.0158$, giving $z = (0.084 - 0.050)/0.0158 \approx 2.15$, consistent with the printout. Since $P = 0.032 < 0.05$, **reject** H_0: there is convincing evidence that the true click-through rates differ between the two banner designs.

6.8.2 Choosing Between a Confidence Interval and a Significance Test

The choice of procedure follows directly from the question being asked. Use a **confidence interval** when the goal is to *estimate* an unknown proportion (or a difference in proportions) — the question takes the form "what is the value of p?" or "how large is the gap between p_1 and p_2?" Use a **significance test** when the goal is to *decide between two competing claims* about a proportion — the question takes the form "is p equal to this specific value, or not?" or "do p_1 and p_2 actually differ?" The two are connected, as shown above: a two-sided test at level α agrees with whether the matching $(1 - \alpha)$ CI contains the hypothesized value. But they answer different questions — a CI reports a plausible range of values for the parameter; a test reports whether one particular value can be ruled out as implausible.

6.9 Common Pitfalls

(!) Lỗi thường gặp: Diễn giải khoảng tin cậy sai là lỗi phổ biến nhất: KHÔNG được nói "có 95% xác suất p nằm trong khoảng này". Cách nói đúng: "chúng ta **tin tưởng** 95% rằng khoảng này chứa giá trị thật của p" — vì p là hằng số cố định, còn khoảng tin cậy mới là cái ngẫu nhiên (thay đổi theo mẫu).

(!) Lỗi thường gặp: Đừng lẫn lộn công thức **sai số chuẩn** (standard error) dùng cho **khoảng tin cậy** với công thức dùng cho **kiểm định giả thuyết**. Với MỘT tỉ lệ: khoảng tin cậy dùng $\hat{p}$ (giá trị ước lượng từ mẫu, vì p thật chưa biết) $\Rightarrow SE = \sqrt{\hat{p}(1 - \hat{p})/n}$; kiểm định dùng p_0 (giá trị giả thuyết trong H_0, vì phép kiểm định *tạm giả sử* H_0 đúng) $\Rightarrow SE = \sqrt{p_0(1 - p_0)/n}$. Với HAI tỉ

lệ tương tự: khoảng tin cậy cho $p_1 - p_2$ dùng công thức **không gộp** (unpooled, mỗi mẫu dùng $\hat{p}$ riêng); kiểm định $H_0 : p_1 = p_2$ dùng công thức **gộp** (pooled) $\hat{p}_c$, vì H_0 giả sử hai tỉ lệ tổng thể bằng nhau nên ước lượng chung bằng cách gộp cả hai mẫu. Dùng nhầm công thức SE là lỗi tính toán rất phổ biến, ngay cả khi các bước còn lại đều đúng.

6.10 Practice problems

Bài tập luyện tập

A city of $850{,}000$ residents is the population of interest. A researcher randomly samples $n = 600$ residents (without replacement) to estimate support for a new transit line; 372 support it. Which of the following is correct?

(A) All three conditions (Random, 10%, Large Counts) for a one-proportion z interval are satisfied

(B) The 10% condition fails

(C) The Large Counts condition fails

(D) The Random condition cannot be verified

Lời giải chi tiết

$\hat{p} = 372/600 = 0.62$. **Random:** SRS stated. **10%:** $0.10(850{,}000) = 85{,}000 \geq 600$. **Large Counts:** $n\hat{p} = 372 \geq 10$, $n(1-\hat{p}) = 228 \geq 10$. All three hold. **(A).**

Bài tập luyện tập

A poll of $n = 40$ people finds $\hat{p} = 0.22$. Is it appropriate to use the one-proportion z-interval here?

(A) Yes, $n = 40$ is large enough

(B) No — $n\hat{p} = 8.8 < 10$, so the Large Counts condition fails

(C) No — the 10% condition fails

(D) Yes, because $\hat{p} < 0.5$

Lời giải chi tiết

$n\hat{p} = 40(0.22) = 8.8 < 10$, so the Normal approximation used to build a z-interval is not justified with this sample; a larger sample (or an adjusted method) is needed. **(B).**

Bài tập luyện tập

A university club has $N = 180$ members. The club president wants to construct a z confidence interval for the proportion who favor moving meetings online and randomly samples $n = 45$ members without replacement, finding 27 in favor. Check all three conditions and state whether the one-proportion z-interval is appropriate.

Lời giải chi tiết

$\hat{p} = 27/45 = 0.6$. **Random:** SRS given — satisfied. **10%:** $0.10(180) = 18$; but $n = 45 > 18$ — **fails**. **Large Counts:** $n\hat{p} = 27 \geq 10$, $n(1-\hat{p}) = 18 \geq 10$ — satisfied on its own. Because the 10% condition fails (more than 10% of a small finite population was sampled without replacement), the standard one-proportion z-interval is **not appropriate** here without adjustment.

Bài tập luyện tập

If the sample size n in a one-proportion confidence interval is quadrupled (multiplied by 4) and $\hat{p}$ stays the same, what happens to the margin of error?

(A) It is multiplied by 4
(B) It is multiplied by 2
(C) It is divided by 2
(D) It is divided by 4

Lời giải chi tiết

$ME = z^* \sqrt{\hat{p}(1-\hat{p})/n}$, so $ME \propto 1/\sqrt{n}$. Quadrupling n divides ME by $\sqrt{4} = 2$. **(C)**.

Bài tập luyện tập

A random sample of $n = 250$ small businesses finds that 145 report using social media for advertising. Construct and interpret a 90% confidence interval for the true proportion of all small businesses that use social media for advertising.

Lời giải chi tiết

$\hat{p} = 145/250 = 0.58$. Conditions (Random, 10%, Large Counts: $n\hat{p} = 145 \geq 10$, $n(1-\hat{p}) = 105 \geq 10$) are satisfied.

$$SE = \sqrt{\frac{0.58(0.42)}{250}} \approx 0.0312, \qquad ME = 1.645(0.0312) \approx 0.0514.$$

90% CI: $0.58 \pm 0.0514 = (0.529,\ 0.631)$. We are 90% confident the true proportion of all small businesses using social media for advertising is between 52.9% and 63.1%.

Bài tập luyện tập

As the confidence level of a one-proportion z-interval increases from 90% to 99% (with n and $\hat{p}$ held fixed), the margin of error:

(A) Decreases
(B) Increases
(C) Stays the same
(D) Cannot be determined without knowing $\hat{p}$

Lời giải chi tiết

A higher confidence level requires a larger critical value (z^* rises from 1.645 to 2.576), which directly increases $ME = z^* \cdot SE$. **(B)**.

Bài tập luyện tập

A city council wants to estimate, with 98% confidence and a margin of error of at most 0.025 (2.5 percentage points), the proportion of residents who support a proposed park renovation. A previous survey suggested about 40% support such projects. Find the required sample size.

Lời giải chi tiết

$z^* = 2.326$ for 98% confidence; $p^* = 0.40$, $1 - p^* = 0.60$.

$$n = \frac{(2.326)^2(0.40)(0.60)}{(0.025)^2} = \frac{5.4103(0.24)}{0.000625} = \frac{1.2985}{0.000625} \approx 2077.5 \;\Rightarrow\; n = 2{,}078.$$

Bài tập luyện tập

A nutritionist samples $n = 90$ adults and finds that 54 meet the recommended daily fiber intake. Construct a 95% confidence interval for the true proportion of adults meeting this recommendation, and write ONE correct sentence interpreting the confidence *level* and ONE correct sentence interpreting the *specific interval* you found.

Lời giải chi tiết

$\hat{p} = 54/90 = 0.6$.

$$SE = \sqrt{\frac{0.6(0.4)}{90}} \approx 0.0516, \qquad ME = 1.96(0.0516) \approx 0.1012.$$

95% CI: $0.6 \pm 0.1012 = (0.499,\ 0.701)$.
Confidence level: if we repeated this sampling process many times and built a 95% CI from each sample, about 95% of those intervals would capture the true population proportion.
Specific interval: we are 95% confident that the interval $(0.499,\ 0.701)$ captures the true proportion of all adults meeting the recommended daily fiber intake.

Bài tập luyện tập

A cereal company claims that 25% of its cereal boxes contain a prize. A consumer group tests $H_0 : p = 0.25$ vs. $H_a : p \neq 0.25$ using a random sample of $n = 300$ boxes, of which 87 contain a prize. Find the test statistic z.

(A) $z \approx 1.60$
(C) $z \approx 2.50$
(B) $z \approx 0.04$
(D) $z \approx 0.16$

Lời giải chi tiết

$\hat{p} = 87/300 = 0.29$. $SE = \sqrt{0.25(0.75)/300} = \sqrt{0.000625} = 0.025$. $z = (0.29 - 0.25)/0.025 = 1.60$. **(A)**.

Bài tập luyện tập

An airline claims that at least 85% of its flights depart on time. A watchdog group suspects the actual on-time rate is lower and tests $H_0 : p = 0.85$ vs. $H_a : p < 0.85$ using a random sample of $n = 150$ flights, of which 116 departed on time. Carry out the significance test at $\alpha = 0.05$ and state a conclusion in context.

Lời giải chi tiết

$\hat{p} = 116/150 \approx 0.7733$. Conditions: Random given; 10% satisfied (150 flights $\ll 10\%$ of all flights); Large Counts using p_0: $np_0 = 127.5 \geq 10$, $n(1-p_0) = 22.5 \geq 10$.

$$SE = \sqrt{\frac{0.85(0.15)}{150}} \approx 0.02916, \qquad z = \frac{0.7733 - 0.85}{0.02916} \approx -2.63.$$

P-value $= P(Z < -2.63) = 1 - \Phi(2.63)$. Interpolating between $2.60 \to 0.9953$ and $2.70 \to 0.9965$: $\Phi(2.63) \approx 0.9957$, so P-value ≈ 0.0043. Since $0.0043 < 0.05$, **reject** H_0: there is convincing evidence the true on-time departure rate is below 85%.

Bài tập luyện tập

In a significance test for a proportion, a P-value of 0.03 means:

(A) There is a 3% probability that H_0 is true

(B) There is a 3% probability that H_a is true

(C) If H_0 is true, there is a 3% probability of getting a sample statistic at least as extreme as the one observed

(D) The true proportion equals the observed sample proportion 3% of the time

Lời giải chi tiết

The P-value is a conditional probability computed *assuming H_0 is true* — it is the probability of a result at least as extreme as what was observed, not the probability that any hypothesis is true. **(C)**.

Bài tập luyện tập

A cell-phone manufacturer claims that no more than 3% of its phones are returned for defects within the first year ($p_0 = 0.03$). A regulator suspects the true return rate is higher and randomly samples $n = 500$ phones sold last year, finding that 22 were returned for defects. Carry out an appropriate test at $\alpha = 0.01$. Use State/Plan/Do/Conclude.

Lời giải chi tiết

State: let p = the true first-year defect-return proportion. $H_0 : p = 0.03$; $H_a : p > 0.03$; $\alpha = 0.01$.
Plan: one-sample z test. Random — given. 10% — 500 phones is a tiny fraction of all phones sold. Large Counts (use p_0): $np_0 = 15 \geq 10$, $n(1-p_0) = 485 \geq 10$ — met.
Do: $\hat{p} = 22/500 = 0.044$.

$$SE = \sqrt{\frac{0.03(0.97)}{500}} \approx 0.00763, \qquad z = \frac{0.044 - 0.03}{0.00763} \approx 1.84.$$

P-value $= P(Z > 1.84) = 1 - \Phi(1.84)$. Interpolating between $1.80 \to 0.9641$ and $1.90 \to 0.9713$: $\Phi(1.84) \approx 0.9670$, so P-value ≈ 0.033.
Conclude: since $0.033 > 0.01$, **fail to reject** H_0. There is not convincing evidence, at the 1% significance level, that the true defect-return rate exceeds 3%.

Bài tập luyện tập

A vending-machine manufacturer claims that exactly 10% of cups are underfilled beyond tolerance ($p_0 = 0.10$). A technician randomly inspects $n = 200$ cups and finds 28 underfilled beyond tolerance. Test whether the true underfill proportion differs from 10% at $\alpha = 0.05$. Use State/Plan/Do/Conclude.

Lời giải chi tiết

State: let $p =$ the true proportion of all cups underfilled beyond tolerance. $H_0 : p = 0.10$; $H_a : p \neq 0.10$; $\alpha = 0.05$.
Plan: one-sample z test. Random — given. 10% — 200 cups is tiny relative to total cups dispensed. Large Counts (p_0): $np_0 = 20 \geq 10$, $n(1-p_0) = 180 \geq 10$ — met.
Do: $\hat{p} = 28/200 = 0.14$.

$$SE = \sqrt{\frac{0.10(0.90)}{200}} \approx 0.02121, \qquad z = \frac{0.14 - 0.10}{0.02121} \approx 1.89.$$

Two-sided P-value $= 2[1 - \Phi(1.89)]$. Interpolating between $1.80 \to 0.9641$ and $1.90 \to 0.9713$: $\Phi(1.89) \approx 0.9706$, so P-value $\approx 2(0.0294) = 0.059$.
Conclude: since $0.059 > 0.05$, **fail to reject** H_0. There is not convincing evidence, at the 5% level, that the true underfill proportion differs from 10% (though the result is close to significant).

Bài tập luyện tập

If a two-sided test of $H_0 : p = p_0$ produces a P-value of 0.08, what would the one-sided P-value be for the alternative in the observed direction (assuming the sample result falls in that direction)?

(A) 0.16

(B) 0.08

(C) 0.04

(D) Cannot be determined from the two-sided P-value alone

Lời giải chi tiết

The two-sided P-value is twice the one-sided P-value in a symmetric Normal test, so the one-sided value is $0.08/2 = 0.04$. **(C)**.

Bài tập luyện tập

In a test of $H_0 : p = 0.5$, a Type I error would mean:

(A) Concluding $p \neq 0.5$ when in fact $p = 0.5$

(B) Concluding $p = 0.5$ when in fact $p \neq 0.5$

(C) Using the wrong sample size

(D) Failing to check conditions

Lời giải chi tiết

A **Type I error** is rejecting a true H_0 — here, concluding the proportion differs from 0.5 when it actually equals 0.5. **(A)**.

Bài tập luyện tập

An airport security screening process uses $H_0 : p = 0.001$ (the normal proportion of bags requiring secondary screening) vs. $H_a : p > 0.001$ (an elevated rate suggesting a system malfunction). Describe, in context, what a Type II error would be for this test and describe one realistic consequence of making it.

Lời giải chi tiết

A **Type II error** here means failing to reject H_0 — concluding there is no evidence of an elevated screening-flag rate — when in fact the true rate *is* elevated (a real malfunction or emerging issue exists). *Consequence:* a genuine security or equipment problem goes undetected and uncorrected, potentially allowing a real risk to persist for longer than necessary before anyone investigates.

Bài tập luyện tập

All else held constant, if the significance level α of a test is increased from 0.01 to 0.10, what happens to the power of the test?

(A) Power decreases

(B) Power increases

(C) Power is unaffected by α

(D) Power becomes exactly $1 - \alpha$

Lời giải chi tiết

Increasing α enlarges the rejection region, making it easier to reject H_0 whenever H_a is true, which raises power — at the cost of also raising the Type I error rate. **(B)**.

Bài tập luyện tập

Two studies test $H_0 : p = 0.40$ vs. $H_a : p > 0.40$ at $\alpha = 0.05$. Suppose the true proportion is actually $p = 0.50$ in both studies. Study 1 uses $n = 60$; Study 2 uses $n = 240$ (four times as large). Without carrying out the full power calculation, explain which study has higher power and why, referring to the standard error.

Lời giải chi tiết

Since $SE = \sqrt{p(1-p)/n}$ shrinks as n grows, Study 2's sampling distributions (both the one used to set the rejection region under H_0, and the one describing actual sample behavior under the true $p = 0.50$) are much narrower than Study 1's. Narrower, more separated distributions mean a much larger share of Study 2's true sampling distribution lies past the critical value in the rejection region. **Study 2 ($n = 240$) has substantially higher power** to detect the same true difference (0.50 vs. 0.40) than Study 1 ($n = 60$).

Bài tập luyện tập

When constructing a confidence interval for $p_1 - p_2$, why is the *unpooled* standard error used instead of a pooled estimate?

(A) Because pooling is mathematically impossible for two samples

(B) Because a confidence interval does not assume $p_1 = p_2$, so each sample's own proportion should describe its own variability

(C) Because the pooled estimate is always larger and therefore more conservative

(D) Because the Large Counts condition cannot be checked with a pooled estimate

Lời giải chi tiết

A confidence interval is used precisely because p_1 and p_2 are unknown and not assumed equal, so each sample's own $\hat{p}$ should describe its own sampling variability. **(B)**.

Bài tập luyện tập

Independent random samples are taken from two high schools to estimate the difference in the proportion of seniors planning to attend a four-year college. School X: $n_1 = 140$ seniors sampled, 98 plan to attend. School Y: $n_2 = 160$ seniors sampled, 88 plan to attend. Construct and interpret a 90% confidence interval for $p_X - p_Y$.

Lời giải chi tiết

$\hat{p}_X = 98/140 = 0.70$; $\hat{p}_Y = 88/160 = 0.55$. Conditions (Random, 10%, independent samples, Large Counts: $98, 42, 88, 72$ all ≥ 10) are satisfied.

$$SE = \sqrt{\frac{0.70(0.30)}{140} + \frac{0.55(0.45)}{160}} = \sqrt{0.0015 + 0.001547} \approx 0.0552.$$

$$ME = 1.645(0.0552) \approx 0.0908. \qquad 90\%\ \text{CI}: \ 0.15 \pm 0.0908 = (0.059,\ 0.241).$$

We are 90% confident that the true difference in proportions planning to attend a four-year college (School X minus School Y) is between 0.059 and 0.241. Since the entire interval is positive, this is convincing evidence that School X's true proportion is higher.

Bài tập luyện tập

A 95% confidence interval for $p_1 - p_2$ is calculated to be $(-0.04,\ 0.11)$. Which conclusion is correctly supported?

(A) We can be sure $p_1 = p_2$

(B) Because the interval contains 0, it is plausible there is no difference between the two population proportions; a two-sided test at $\alpha = 0.05$ would fail to reject $H_0 : p_1 = p_2$

(C) Because the interval is mostly positive, we can conclude $p_1 > p_2$

(D) The interval is invalid because it contains a negative value

Lời giải chi tiết

An interval containing 0 means a difference of 0 is a plausible value for $p_1 - p_2$; by the CI–test connection, the corresponding two-sided test at $\alpha = 0.05$ would fail to reject $H_0 : p_1 = p_2$. **(B)**.

Bài tập luyện tập

An online retailer redesigns its checkout page. Old page: 30 of 250 visitors converted. New page: 51 of 300 visitors converted. Test $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$ at $\alpha = 0.05$ (1 =old, 2 =new).

(A) $z \approx 1.65$, $P \approx 0.10$; fail to reject H_0 at $\alpha = 0.05$

(B) $z \approx 1.65$, $P \approx 0.05$; reject H_0

(C) $z \approx 3.30$, $P \approx 0.001$; reject H_0

(D) Cannot test; sample sizes differ

Lời giải chi tiết

$\hat{p}_1 = 30/250 = 0.12$, $\hat{p}_2 = 51/300 = 0.17$. Pooled: $\hat{p}_c = \dfrac{30+51}{250+300} = \dfrac{81}{550} \approx 0.1473$.

$$SE = \sqrt{0.1473(0.8527)\left(\tfrac{1}{250} + \tfrac{1}{300}\right)} = \sqrt{0.1256(0.007333)} \approx \sqrt{0.000921} \approx 0.0304.$$

$z = \dfrac{0.17 - 0.12}{0.0304} \approx 1.65$, two-sided $P \approx 0.10$. Since $0.10 > 0.05$, **fail to reject** H_0 — not quite convincing evidence of a difference at the 5% level (though suggestive). **(A)**.

Bài tập luyện tập

Sample 1: $x_1 = 45$ successes in $n_1 = 150$. Sample 2: $x_2 = 60$ successes in $n_2 = 150$. Find the pooled sample proportion $\hat{p}_c$ used for a two-proportion significance test.

(A) 0.30

(B) 0.35

(C) 0.40

(D) 0.175

Lời giải chi tiết

$\hat{p}_c = \dfrac{45+60}{150+150} = \dfrac{105}{300} = 0.35$. **(B)**.

Bài tập luyện tập

A university offers a traditional lecture section and a "flipped classroom" section of introductory statistics. Of $n_1 = 90$ students in the flipped section, 72 earned a grade of B or higher. Of $n_2 = 110$ students in the traditional section, 71 earned a grade of B or higher. Test whether the true proportion earning a B or higher is greater in the flipped section, at $\alpha = 0.05$. Use State/Plan/Do/Conclude.

Lời giải chi tiết

State: let p_1, p_2 = true proportions earning B or higher in the flipped and traditional sections. $H_0 : p_1 = p_2$; $H_a : p_1 > p_2$; $\alpha = 0.05$.
Plan: two-sample z test (pooled). Random/independent groups assumed; 10% satisfied; pooled $\hat{p}_c = \dfrac{72+71}{90+110} = \dfrac{143}{200} = 0.715$; Large Counts: $n_1\hat{p}_c \approx 64.4$, $n_1(1-\hat{p}_c) \approx 25.7$, $n_2\hat{p}_c \approx 78.7$, $n_2(1-\hat{p}_c) \approx 31.4$, all ≥ 10 — met.

Do: $\hat{p}_1 = 72/90 = 0.80$; $\hat{p}_2 = 71/110 \approx 0.6455$.

$$SE = \sqrt{0.715(0.285)\left(\tfrac{1}{90} + \tfrac{1}{110}\right)} = \sqrt{0.203775(0.020202)} \approx 0.0642.$$

$$z = \frac{0.80 - 0.6455}{0.0642} \approx 2.41.$$

P-value $= 1 - \Phi(2.41)$. Interpolating between $2.40 \to 0.9918$ and $2.50 \to 0.9938$: $\Phi(2.41) \approx 0.9920$, so P-value ≈ 0.0080.
Conclude: since $0.0080 < 0.05$, **reject** H_0. There is convincing evidence that the true proportion earning a B or higher is greater in the flipped-classroom section.

Bài tập luyện tập

Pooling the two sample proportions into a single $\hat{p}_c$ is appropriate for:

(A) Both the confidence interval and the significance test for $p_1 - p_2$

(B) The confidence interval only, since it estimates the common variability under H_0

(C) The significance test only, since H_0 assumes $p_1 = p_2$, so it is estimated using both samples combined

(D) Neither procedure; pooling is only used for one-proportion procedures

Lời giải chi tiết

Only the significance test assumes $p_1 = p_2$ under H_0; the confidence interval makes no such assumption and always uses the unpooled SE. **(C)**.

Bài tập luyện tập

A researcher wants to know only whether the proportion of defective items has changed from a historical value of 8%; she does not need to estimate the new proportion's actual value. Which procedure is most directly suited to her goal?

(A) A one-proportion confidence interval

(B) A one-proportion significance test

(C) A two-proportion confidence interval

(D) A two-proportion significance test

Lời giải chi tiết

Testing whether a proportion has changed from a specific historical value is exactly what a significance test (comparing $\hat{p}$ to a hypothesized p_0) is designed to answer. **(B)**.

Bài tập luyện tập

Explain, in one or two sentences each: (a) why the standard error formula changes between a one-proportion confidence interval and a one-proportion significance test; (b) why increasing the sample size decreases the margin of error of a confidence interval *and* increases the power of a significance test testing a fixed true effect.

Lời giải chi tiết

(a) The CI's SE uses $\hat{p}$ because the true p is unknown and is exactly what is being estimated; the test's SE uses the hypothesized p_0 because a test temporarily assumes H_0 is true in order to measure how far the observed statistic is from what H_0 predicts.
(b) Both effects trace back to the same source: $SE = \sqrt{p(1-p)/n}$ shrinks as n grows. A smaller SE directly shrinks the margin of error, producing a tighter CI. A smaller SE also makes the sampling distributions under H_0 and under a fixed true alternative narrower and therefore less overlapping, so a larger share of the alternative's distribution falls past the critical value into the rejection region — i.e., power increases.

Bài tập luyện tập

A city's building code review board has historically approved 75% of routine renovation permit applications ($p_0 = 0.75$). After adopting a new streamlined review process, the board reviews a random sample of $n = 180$ recent applications and approves 150 of them. (a) Test whether the true approval rate under the new process differs from 75%, at $\alpha = 0.05$, using State/Plan/-Do/Conclude. (b) If the board incorrectly concludes the approval rate has changed when it actually has not, what type of error is this, and what is one realistic consequence?

Lời giải chi tiết

(a) State: let $p =$ the true approval proportion under the new process. $H_0 : p = 0.75$; $H_a : p \neq 0.75$; $\alpha = 0.05$.
Plan: one-sample z test. Random — given. 10% — reasonable to treat the ongoing stream of applications as effectively unlimited. Large Counts (p_0): $np_0 = 135 \geq 10$, $n(1-p_0) = 45 \geq 10$ — met.
Do: $\hat{p} = 150/180 \approx 0.8333$.

$$SE = \sqrt{\frac{0.75(0.25)}{180}} \approx 0.03228, \qquad z = \frac{0.8333 - 0.75}{0.03228} \approx 2.58.$$

Two-sided P-value $= 2[1 - \Phi(2.58)]$. Interpolating between $2.55 \to 0.9946$ and $2.60 \to 0.9953$: $\Phi(2.58) \approx 0.9950$, so P-value $\approx 2(0.0050) = 0.010$.
Conclude: since $0.010 < 0.05$, **reject** H_0. There is convincing evidence that the true approval rate under the new process differs from 75% (the sample suggests it is now higher).

(b) Concluding the rate has changed when it actually has *not* is a **Type I error** (rejecting a true H_0). *Consequence:* the board might unnecessarily overhaul or further modify a review process that was never actually broken, wasting staff time and resources for no real benefit.

Bài tập luyện tập

A very large clinical trial finds a statistically significant difference ($P = 0.001$) between the cure rates of two treatments, but the actual difference in cure rates is only 0.3 percentage points. What can correctly be concluded?

(A) The treatments are clearly meaningfully different in practice

(B) The result is statistically significant, but the tiny effect size means the difference may not be practically important

(C) Because $P < 0.05$, the sample size must have been too small

(D) The test is invalid whenever the effect size is small

Lời giải chi tiết

With a very large sample, even a trivial, practically unimportant difference can be statistically significant, because the standard error shrinks as n grows. Statistical significance says the difference is unlikely to be due to chance alone; it says nothing about whether the difference is large enough to matter in practice. **(B)**.

Bài tập luyện tập

A significance test of $H_0 : p = 0.40$ produces a P-value of 0.21. Explain precisely what this P-value does and does not tell us.

Lời giải chi tiết

It tells us that, *if* H_0 were true ($p = 0.40$), there would be about a 21% chance of observing a sample proportion at least as far from 0.40 (in the direction of H_a) as the one actually obtained — an unremarkable, unsurprising result under H_0, so there is no convincing evidence against it. It does **not** tell us the probability that H_0 is true, the probability that H_a is true, or that H_0 is exactly correct — a P-value only measures how surprising the observed sample result would be if H_0 held.

Bài tập luyện tập

Two independent samples give $\hat{p}_1 = 0.45$ ($n_1 = 80$) and $\hat{p}_2 = 0.30$ ($n_2 = 100$). Find the (unpooled) standard error used for a confidence interval for $p_1 - p_2$.

Lời giải chi tiết

$$SE = \sqrt{\frac{0.45(0.55)}{80} + \frac{0.30(0.70)}{100}} = \sqrt{0.003094 + 0.0021} = \sqrt{0.005194} \approx 0.0721.$$

Bài tập luyện tập

A quality engineer wants to test a claim about the proportion of defective bolts in a batch of exactly $N = 60$ bolts, and inspects *all* 60 bolts (a complete census, not a sample). Explain why standard one-proportion z-procedures are not appropriate here, regardless of how large $n\hat{p}$ and $n(1-\hat{p})$ turn out to be.

Lời giải chi tiết

z-procedures for a proportion exist to quantify the uncertainty of estimating an *unknown* population proportion from a *sample* that is only part of the population. Here, every single bolt in the population of interest was inspected, so there is no sampling variability at all: the "sample" proportion *is* the population proportion exactly, with nothing left to estimate or generalize beyond what was already completely measured. A confidence interval or significance test is only meaningful when a subset of a larger population is used to infer about that population — neither applies to a full census.

Bài tập luyện tập

A coin is tested for fairness: $H_0 : p = 0.5$ vs. $H_a : p \neq 0.5$, based on $n = 400$ flips resulting in 236 heads. Test at $\alpha = 0.05$.

Lời giải chi tiết

$\hat{p} = 236/400 = 0.59$.

$$SE = \sqrt{\frac{0.5(0.5)}{400}} = \sqrt{0.000625} = 0.025, \qquad z = \frac{0.59 - 0.50}{0.025} = 3.60.$$

This z-value lies beyond the largest entry in the standard Normal table ($3.00 \to 0.9987$), so the two-sided P-value is smaller than $2(1 - 0.9987) = 0.0026$ — in fact it is far smaller still ($P < 0.001$). Since this is far below $\alpha = 0.05$, **reject** H_0: there is very convincing evidence the coin is not fair (it favors heads).

Bài tập luyện tập

In a two-proportion z-test with unequal sample sizes $n_1 \neq n_2$, the pooled proportion $\hat{p}_c = \dfrac{x_1 + x_2}{n_1 + n_2}$ is best described as:

(A) Always exactly the simple average of $\hat{p}_1$ and $\hat{p}_2$

(B) A weighted average of $\hat{p}_1$ and $\hat{p}_2$, weighted by each sample's size

(C) Always equal to the smaller of $\hat{p}_1$ and $\hat{p}_2$

(D) Undefined whenever $n_1 \neq n_2$

Lời giải chi tiết

Writing $x_1 = n_1\hat{p}_1$ and $x_2 = n_2\hat{p}_2$, the pooled proportion becomes $\hat{p}_c = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ — a weighted average of the two sample proportions, with the larger sample contributing more weight. **(B)**.

Bài tập luyện tập

If the planning value p^* used in the sample-size formula $n = (z^*/ME)^2 p^*(1 - p^*)$ moves from $p^* = 0.5$ to $p^* = 0.1$, with z^* and ME held fixed, the required sample size n will:

(A) Increase

(B) Decrease

(C) Stay exactly the same

(D) Become negative

Lời giải chi tiết

$p^*(1 - p^*)$ is maximized at $p^* = 0.5$ (value 0.25) and is smaller at $p^* = 0.1$ (value 0.09). Since n is directly proportional to $p^*(1 - p^*)$, a smaller product means a smaller required sample size. **(B)**.

Bài tập luyện tập

A poll estimates $\hat{p} = 0.47$ from $n = 500$ respondents and reports a margin of error of about 0.044 at 95% confidence. Verify this margin of error and explain, in one sentence, what it means in context.

Lời giải chi tiết

$$SE = \sqrt{\frac{0.47(0.53)}{500}} \approx 0.0223, \qquad ME = 1.96(0.0223) \approx 0.044.$$

This confirms the reported value. In context: we can be 95% confident that the true population proportion is within about 4.4 percentage points of the sample estimate of 47% — that is, the true proportion is estimated to lie roughly between 42.6% and 51.4%.

Bài tập luyện tập

For a one-proportion significance test, the Large Counts condition should be checked using:

(A) $\hat{p}$, always

(B) p_0, the hypothesized value

(C) The average of $\hat{p}$ and p_0

(D) Large Counts does not apply to significance tests

Lời giải chi tiết

A significance test temporarily assumes H_0 is true, so every part of the test — including checking Large Counts — uses the hypothesized value p_0, not the sample value $\hat{p}$. **(B)**.

Bài tập luyện tập

Two independent surveys ask residents about support for a local ordinance. Survey 1: $n_1 = 90$, 58 in favor. Survey 2: $n_2 = 100$, 55 in favor. Test $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$ at $\alpha = 0.05$.

Lời giải chi tiết

$\hat{p}_1 = 58/90 \approx 0.6444$; $\hat{p}_2 = 55/100 = 0.55$. Pooled: $\hat{p}_c = \dfrac{58+55}{90+100} = \dfrac{113}{190} \approx 0.5947$.

$$SE = \sqrt{0.5947(0.4053)\left(\tfrac{1}{90} + \tfrac{1}{100}\right)} = \sqrt{0.2410(0.02111)} \approx \sqrt{0.005089} \approx 0.0713.$$

$$z = \frac{0.6444 - 0.55}{0.0713} \approx 1.32.$$

Two-sided P-value $= 2[1 - \Phi(1.32)]$. Interpolating between $1.30 \to 0.9032$ and $1.40 \to 0.9192$: $\Phi(1.32) \approx 0.9064$, so P-value $\approx 2(0.0936) = 0.187$. Since $0.187 > 0.05$, **fail to reject** H_0: there is not convincing evidence that true support for the ordinance differs between the two survey populations.

Bài tập luyện tập

A clinic measures the same 150 patients' blood-pressure-control status (controlled/not controlled) before starting a new medication and again 3 months after starting it. Which of the following is true?

(A) A two-sample z test/interval for $p_1 - p_2$ (independent samples) is directly appropriate here

(B) This is a matched-pairs (before/after) design, so the independent two-sample z-procedures in this unit do not directly apply

(C) The 10% condition automatically fails in every matched design

(D) Matched designs always require a larger sample size than independent designs

Lời giải chi tiết

The "before" and "after" measurements come from the very same 150 patients rather than two independent groups, so the observations are paired/correlated. Using the independent two-sample procedures here would ignore that correlation and is not directly appropriate; a matched-pairs approach is needed instead. **(B)**.

Bài tập luyện tập

Explain, in one sentence, why applying the independent two-sample z-test for $p_1 - p_2$ to the before/after study on the same 150 patients (from the previous problem) would be inappropriate.

Lời giải chi tiết

Because the "before" and "after" measurements come from the very same 150 patients rather than from two independent groups, the two sets of observations are correlated (an individual patient's "after" outcome is linked to their own "before" outcome), which violates the independence condition required by the two-sample z procedures and would generally distort the standard error and produce misleading conclusions.

Bài tập luyện tập

A national health agency wants to determine whether the true proportion of a country's adults who are vaccinated exceeds a target of 70%. It separately also wants to estimate, as precisely as possible, the actual current vaccination proportion. Which pairing of procedures is correct for these two separate goals?

(A) Confidence interval for the first goal, significance test for the second

(B) Significance test for the first goal ($H_0 : p = 0.70$ vs. $H_a : p > 0.70$), confidence interval for the second

(C) Two confidence intervals

(D) Two significance tests

Lời giải chi tiết

The first goal ("does p exceed a specific target value?") is a decision between competing claims — a significance test. The second goal ("what is the actual value of p?") is estimation — a confidence interval. **(B)**.

Inference for Quantitative Data: Means

Mục tiêu Unit: Phân phối t và lý do dùng t thay cho z khi σ chưa biết; điều kiện Random – 10% – Normal/Large Sample và tính robust của thủ tục t; khoảng tin cậy và kiểm định t cho một trung bình μ; cho hiệu hai trung bình độc lập $\mu_1 - \mu_2$; cho dữ liệu bắt cặp (matched pairs) trên hiệu số μ_d; và ảnh hưởng của cỡ mẫu lên sai số biên (margin of error).

7.1 Reviewing the Four-Step Framework for Means

All quantitative inference in this unit – confidence intervals and significance tests alike – follows the same four-step framework used for proportions: **State**, **Plan**, **Do**, **Conclude**. The parameter changes from a proportion p to a mean μ (or a difference of means $\mu_1 - \mu_2$, or a mean difference μ_d), and the sampling distribution changes from Normal/z-based to t-based, but the logical structure of a complete inference argument does not change. Two patterns recur throughout every section that follows: a confidence interval always has the form *statistic* $\pm$ *(critical value)*$\times$*(standard error)*, and a test statistic always has the form $\dfrac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$. Only the specific statistic, standard error formula, and degrees of freedom change from one procedure to the next.

The four-step framework, applied to means

State: Define the parameter in context (μ, $\mu_1 - \mu_2$, or μ_d). For a test, write H_0 and H_a and choose a significance level α.

Plan: Name the correct procedure (one-sample t, two-sample t, or matched-pairs t) and check its conditions: **R**andom, **10%**, **N**ormal/Large Sample.

Do: If conditions are met, compute the standard error, the test statistic or margin of error, the degrees of freedom, and the P-value (or confidence interval).

Conclude: Interpret the result in the context of the original question – reject or fail to reject H_0 in words, or state what the interval says about the parameter.

VN **GIẢI THÍCH TIẾNG VIỆT**

Khung **4 bước** (State – Plan – Do – Conclude) đã xuất hiện xuyên suốt phần suy diễn cho tỉ lệ và được áp dụng lại nguyên vẹn cho **trung bình** ở Unit này. Điểm khác biệt cốt lõi: khi σ (độ lệch chuẩn tổng thể) không biết – gần như luôn luôn trong thực tế – ta phải ước lượng nó bằng s (độ lệch chuẩn mẫu), và điều này buộc ta chuyển từ phân phối chuẩn (z) sang phân phối t. Nhiều ví dụ mẫu trong chương này sẽ trình bày rõ 4 bước bằng các nhãn in đậm để rèn thói quen trình bày bài làm FRQ đúng chuẩn.

7.2 The t-Distribution

Properties of the t-distribution

- Symmetric and unimodal, centered at 0 – just like the standard Normal curve.
- Indexed by degrees of freedom df: a different t-curve exists for every value of df.
- More spread out than the standard Normal curve (lower peak, heavier tails), because it must account for the extra uncertainty of estimating σ with s.
- As $df \to \infty$, the t-distribution converges to the standard Normal distribution.
- The total area under every t-curve equals 1, exactly as for any density curve.

VN **GIẢI THÍCH TIẾNG VIỆT**

Ghi nhớ 5 tính chất của phân phối t: đối xứng quanh 0, có tham số df riêng cho từng đường cong, tản mạn hơn (đỉnh thấp, đuôi dày) so với z, hội tụ về z khi $df \to \infty$, và luôn có tổng diện tích bằng 1 như mọi đường cong mật độ khác.

When σ is unknown (the usual case), we estimate it with s and use the t-distribution with $df = n - 1$:

$$\text{CI: } \bar{x} \pm t^* \frac{s}{\sqrt{n}}, \qquad \text{Test: } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

Conditions: Random; 10%; and Normal/Large Sample – population approximately Normal, *or* $n \geq 30$, *or* the sample shows no strong skew/outliers for smaller n.

7.2.1 Why We Use t Instead of z

For a sample mean $\bar{x}$, the true standardized statistic $\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ follows a standard Normal distribution exactly, *provided* σ is known. In practice σ is almost never known – we only have the data in front of us – so we substitute the sample standard deviation s in its place, forming

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

Because s itself varies from sample to sample (it is a statistic, not a fixed constant), this substitution injects an extra source of random variability beyond what $\bar{x}$ alone contributes. The resulting statistic no longer follows a standard Normal distribution: it follows a slightly different, more spread-out distribution called the **t-distribution**, first derived by William Sealy Gosset (writing under the pseudonym "Student") in 1908 while working on quality control at the Guinness brewery.

This extra spread matters most when n is small: with only a handful of observations, s can land quite far from the true σ just by chance, so the t-distribution must have noticeably heavier tails to keep the stated confidence level accurate. As the sample grows, s becomes a far more reliable estimate of σ, and the t-distribution's extra spread shrinks toward nothing.

Substituting the estimate s for the unknown σ adds extra sampling variability. The t-distribution accounts for this by being **more spread out** than the standard Normal distribution, especially for small sample sizes (where s itself is a noisier estimate of σ).

VN **GIẢI THÍCH TIẾNG VIỆT**

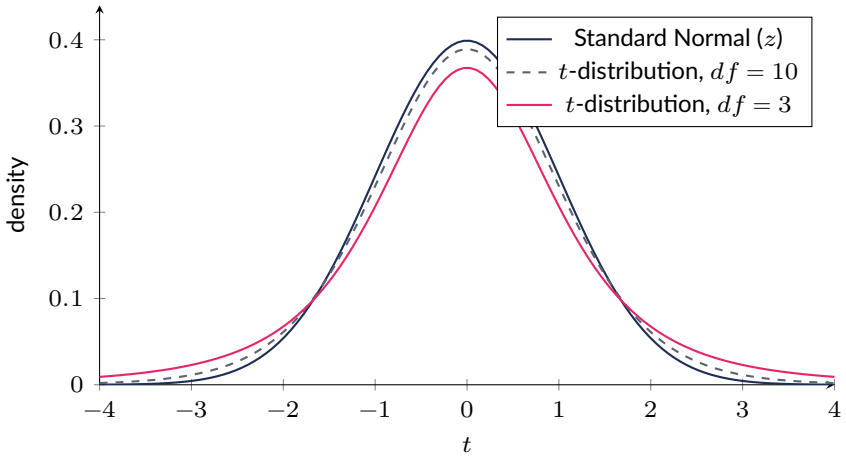
Lý do dùng **phân phối** t thay vì z: công thức chuẩn hoá $\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ chỉ thực sự có phân phối chuẩn

khi σ đã biết. Vì σ hầu như luôn *không biết* trong thực tế, ta phải thay bằng s ước lượng từ mẫu – và s lại dao động ngẫu nhiên theo từng mẫu khác nhau. Sự bất định thêm vào này khiến thống kê $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ có phân phối **tản mạn hơn** (đuôi dày hơn) so với phân phối chuẩn – đặc biệt rõ khi cỡ mẫu n nhỏ.

7.2.2 Shape of the t-Distribution and Degrees of Freedom

Every t-distribution is symmetric and bell-shaped, centered at 0, just like the standard Normal curve – but it has **heavier tails** (more probability far from the center) and a **lower, wider peak**. The exact shape is indexed by a single parameter, the **degrees of freedom** df. For one-sample procedures, $df = n - 1$: once the sample mean $\bar{x}$ is fixed, only $n - 1$ of the n deviations $(x_i - \bar{x})$ are free to vary (the last one is determined by the requirement that the deviations sum to zero), so s is built from $n - 1$ independent pieces of information.

As df increases (larger samples), s becomes a more precise, more stable estimate of σ, so the extra variability shrinks and the t-distribution's shape converges toward the standard Normal curve. In the limit $df \to \infty$, the t-distribution *is* the standard Normal distribution – which is exactly why the $df = \infty$ row of a t-table reproduces familiar z^* critical values (1.645, 1.960, 2.326, ...).



The t-distribution with $df = 3$ (pink) is noticeably lower-peaked and heavier-tailed than the standard Normal curve (navy); with $df = 10$ (dashed grey) it has already moved much closer to Normal. As $df \to \infty$, the t-curve converges to the standard Normal curve.

VN **GIẢI THÍCH TIẾNG VIỆT**

Phân phối t luôn **đối xứng, hình chuông**, tâm tại 0 – giống phân phối chuẩn – nhưng **đuôi dày hơn** và **đỉnh thấp hơn**. Hình dạng cụ thể phụ thuộc vào **bậc tự do** $df = n - 1$ (với suy diễn một mẫu). n càng lớn $\Rightarrow$ df càng lớn $\Rightarrow$ s ước lượng σ càng chính xác $\Rightarrow$ đường cong t càng "áp sát" đường cong chuẩn z. Khi $df \to \infty$, phân phối t trở thành chính phân phối chuẩn – đó là lý do hàng $df = \infty$ trong bảng t trùng với các giá trị z^* quen thuộc.

7.2.3 Reading Critical Values from the t-Table

A t-table lists critical values t^* for a grid of degrees of freedom (rows) and one-tail (upper-tail) areas (columns). To find t^* for a confidence interval, first compute $df = n - 1$, then find the column whose area equals $\dfrac{1 - C}{2}$, where C is the confidence level (this splits the leftover probability $1 - C$ evenly between the two tails).

Reading the t-table

(a) A researcher plans a 98% confidence interval for μ from a sample of $n = 18$. Find t^*. $df = n - 1 = 17$. A 98% two-sided interval leaves $1 - 0.98 = 0.02$ total in the tails, so each tail has area 0.01. Reading the $df = 17$ row at the 0.01 column: $t^* = 2.567$. (For comparison, the corresponding z^* for 98% confidence, found on the $df = \infty$ row, is only 2.326 – the t-based value is larger because the t-distribution is more spread out.)

(b) Track how the 95% critical value (t^* at one-tail area 0.025) changes as df increases:

df	5	10	20	30	40	($df = \infty$: z^*)
t^*	2.571	2.228	2.086	2.042	2.021	(1.960)

The values decrease steadily toward 1.960 as $df \to \infty$, exactly the convergence shown in the figure above.

VN **GIẢI THÍCH TIẾNG VIỆT**

Cách đọc **bảng** t: tính $df = n - 1$, xác định diện tích đuôi cần dùng ($= \frac{1-C}{2}$ cho khoảng tin cậy hai phía mức C), rồi dò giao điểm hàng df – cột diện tích đuôi để lấy t^*. Luôn nhớ $t^* > z^*$ ở cùng mức tin cậy (vì t tản mạn hơn), và khoảng cách này thu hẹp dần khi df tăng.

The choice between a z procedure and a t procedure comes down to a single question: is σ known?

	z **procedure (σ known)**	t **procedure (σ unknown)**
Standard error	$\sigma/\sqrt{n}$ (exact)	$s/\sqrt{n}$ (estimated)
Critical value	z^*, standard Normal table	t^*, t-table with $df = n - 1$
Test statistic	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Frequency of use	Rare – only when σ is explicitly given	Nearly always, since σ is virtually never known

7.3 Conditions for One-Sample t Procedures

7.3.1 Random, 10%, and Normal/Large Sample

Every one-sample t procedure – confidence interval or significance test – requires the same three conditions before the mechanics can be trusted. Note the naming: procedures for a *proportion* check a **Large Counts** condition ($np_0 \geq 10$ and $n(1 - p_0) \geq 10$), while procedures for a *mean* check a **Normal/Large Sample** condition instead – the two names are not interchangeable, and mixing them up on a free-response answer costs a scoring point even when the surrounding work is correct.

Random: the data come from a random sample or a randomized experiment, so the sample is representative and the inference is not systematically biased.

10% condition: if sampling without replacement from a finite population, the sample size n should be no more than 10% of the population size, so that individual observations are close enough to independent for the standard error formula $s/\sqrt{n}$ to be valid.

Normal/Large Sample condition: the sampling distribution of $\bar{x}$ must be approximately Normal. This holds if *any one* of the following is true: (i) the population itself is known to be approximately Normal; (ii) $n \geq 30$, so the Central Limit Theorem guarantees $\bar{x}$ is approximately Normal regardless of the population's shape; or (iii) $n < 30$ but a graph of the sample data (dotplot, boxplot, stemplot) shows no strong skew and no outliers.

Each condition guards against a specific failure. **Random** protects against selection bias: a convenience sample of the loudest customers, the easiest-to-reach patients, or the first n items off a production line is not guaranteed to represent the population, no matter how large n is. The **10%** condition exists because the formula $SE = s/\sqrt{n}$ silently assumes the observations behave independently; sampling more than roughly 10% of a finite population without replacement measurably reduces the true variability below what an independence-based formula predicts, so the stated confidence level would no longer be accurate. The **Normal/Large Sample** condition exists because the entire t^*/P-value machinery is built on the assumption that $\bar{x}$ itself has an approximately Normal sampling distribution – without that, the critical values from the t-table simply do not apply.

VN **GIẢI THÍCH TIẾNG VIỆT**

Ba điều kiện bắt buộc trước khi dùng thủ tục t một mẫu: **Random** (mẫu ngẫu nhiên/thí nghiệm ngẫu nhiên hoá), **10%** (cỡ mẫu không vượt quá 10% tổng thể khi lấy mẫu không hoàn lại), và **Normal/Large Sample** (tổng thể xấp xỉ chuẩn, HOẶC $n \geq 30$ nhờ Định lý giới hạn trung tâm, HOẶC $n < 30$ nhưng dữ liệu mẫu không có lệch mạnh/ngoại lai rõ rệt khi vẽ đồ thị).

7.3.2 Robustness of t Procedures

A procedure is called **robust** against a particular violation of its conditions if the actual behavior (true confidence level, true Type I error rate) stays close to the stated level even when that condition is not perfectly met. t-procedures are reasonably robust against *mild* departures from Normality, especially as n grows – but robustness has limits.

> t-procedures are fairly robust to non-Normality **except** when the sample is small *and* shows strong skewness or a clear outlier. As a rough guide: for $n < 15$, only proceed if the data show no outliers and no strong skew; for $15 \leq n < 30$, mild skew is tolerable but strong skew or outliers are still a concern; for $n \geq 30$, the Central Limit Theorem makes t-procedures robust to all but the most extreme outliers.

When checking a small sample by eye, the familiar $1.5 \times IQR$ rule from univariate data analysis remains the standard benchmark for flagging a possible outlier (a value more than $1.5\,IQR$ below Q_1 or above Q_3). Skewness that is severe enough to worry about typically shows up as a visibly long tail on one side of a dotplot or boxplot, with the bulk of the observations bunched toward the other side – not merely a slight asymmetry.

VN **GIẢI THÍCH TIẾNG VIỆT**

Khi kiểm tra dữ liệu mẫu nhỏ bằng mắt, quy tắc $1.5 \times IQR$ quen thuộc từ phần thống kê mô tả một biến vẫn là chuẩn để xác định ngoại lai (một giá trị thấp hơn Q_1 hoặc cao hơn Q_3 quá $1.5\,IQR$). Lệch (skew) đủ nghiêm trọng để đáng lo ngại thường thể hiện rõ bằng một "đuôi" kéo dài hẳn về một phía trên dotplot/boxplot, với phần lớn giá trị dồn về phía còn lại – chứ không chỉ là bất đối xứng nhẹ.

Checking conditions in practice

(a) A class of $n = 34$ students records the time (minutes) needed to finish a survey. A histogram is unimodal and slightly right-skewed with no extreme outliers. Can a one-sample t procedure proceed?

Yes. Even though the population shape is skewed, $n = 34 \geq 30$, so the Large Sample condition is satisfied through the Central Limit Theorem – the sampling distribution of $\bar{x}$ will be approximately Normal regardless of the (mild) skew in the raw data.

(b) A different class of $n = 9$ students has survey times that are mostly 10–15 minutes, except for one student who took 55 minutes. Can a one-sample t procedure proceed?

Not safely. With $n = 9 < 15$ and one extreme outlier, the Normal/Large Sample condition fails: the sample is far too small for the Central Limit Theorem to help, and the outlier strongly suggests the population is not Normal (or at least that $\bar{x}$ and s from this sample are unreliable). One should investigate the outlier, consider removing it only with strong justification, or use a resistant/nonparametric alternative instead of a standard t procedure.

(!) Lỗi thường gặp: Một hiểu lầm phổ biến: "$n \geq 30$ **thì luôn an toàn, bất kể dữ liệu trông như thế nào.**" Định lý giới hạn trung tâm giúp $\bar{x}$ xấp xỉ chuẩn khi n lớn, nhưng nó KHÔNG miễn nhiễm hoàn toàn với các **outlier cực đoan** – một giá trị bất thường quá lớn vẫn có thể kéo lệch $\bar{x}$ và thổi phồng s, dù $n = 30$ hay $n = 300$. Luôn nhìn đồ thị dữ liệu trước khi kết luận điều kiện Normal/Large Sample được thoả mãn.

7.4 The One-Sample t Confidence Interval for μ

When conditions are met, a confidence interval for a population mean μ takes the general form *statistic* $\pm$ *margin of error*:

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}, \qquad df = n - 1.$$

The standard error $SE = \dfrac{s}{\sqrt{n}}$ measures the typical distance between $\bar{x}$ and μ; multiplying by the critical value t^* builds in enough of a cushion that the method captures μ at the stated confidence level in the long run.

Confidence level vs. confidence interval

These are two different objects, and free-response scoring distinguishes them sharply. The **confidence interval** is the specific numeric range computed from one sample (e.g., $(833.5, 866.5)$ hours) – it either does or does not contain the true μ, though we never know which. The **confidence level** (e.g., 95%) describes the *method*: if the same sampling procedure were repeated many times, about 95% of the resulting intervals would capture the true μ. The correct interpretation template is: *"We are $C\%$ confident that the interval (___, ___) captures the true [parameter, described in context]."* Two answers that lose credit: describing individual values instead of the parameter ("$C\%$ of bulbs last between..."), and assigning a probability to the fixed, already-determined parameter ("there is a $C\%$ probability that μ is in this interval").

VN **GIẢI THÍCH TIẾNG VIỆT**

Phân biệt **mức tin cậy** (confidence level, ví dụ 95%) và **khoảng tin cậy** (confidence interval, ví dụ $(833.5, 866.5)$ giờ): mức tin cậy mô tả *phương pháp* – nếu lặp lại việc lấy mẫu nhiều lần, khoảng 95% số khoảng tính được sẽ chứa μ thật; còn khoảng tin cậy là kết quả CỤ THỂ từ MỘT mẫu duy nhất – nó hoặc chứa μ, hoặc không, ta không bao giờ biết chắc. Câu trả lời chuẩn luôn theo mẫu: "Chúng ta tin tưởng $C\%$ rằng khoảng $(...)$ chứa giá trị thật của [tham số, viết theo ngữ cảnh]."

Ví dụ mẫu · A one-sample t interval

A random sample of $n = 25$ light bulbs has $\bar{x} = 850$ hours, $s = 40$ hours. Construct a 95% CI for the true mean lifetime μ (population roughly symmetric, no outliers).
State: parameter of interest is μ, the true mean lifetime of this bulb model. We want a 95% confidence interval.

Plan: one-sample t interval. Random: stated. 10%: 25 bulbs is surely less than 10% of all bulbs of this model produced. Normal/Large Sample: population is described as roughly symmetric with no outliers, so this is acceptable even though $n < 30$.
Do: $df = 24$, $t^* = 2.064$ (from the t-table, $df = 24$, one-tail area 0.025). $SE = \dfrac{40}{\sqrt{25}} = 8$.

$$850 \pm 2.064(8) = 850 \pm 16.51 = (833.5,\ 866.5) \text{ hours.}$$

Conclude: We are 95% confident that the true mean lifetime of this bulb model is between 833.5 and 866.5 hours.

Ví dụ mẫu · A small-sample t interval

A manufacturer records the drying time (minutes) of $n = 12$ randomly selected samples of a new paint formulation:

$$14,\ 15,\ 15,\ 16,\ 16,\ 16,\ 17,\ 17,\ 17,\ 18,\ 18,\ 19.$$

Construct a 90% confidence interval for the true mean drying time μ.
State: parameter is μ, the true mean drying time of the new formulation. Target: 90% CI.
Plan: one-sample t interval, $df = n - 1 = 11$. Random: samples are stated to be randomly selected. 10%: satisfied (batch size is surely far larger than 120 cans). Normal/Large Sample: $n = 12 < 30$, so we examine the dotplot below.

Drying times (minutes) for $n = 12$ paint samples. The distribution is roughly symmetric and single-peaked with no outliers, so the Normal/Large Sample condition is satisfied despite the small sample size.

The dotplot is roughly symmetric, unimodal, and has no outliers, so we may proceed.
Do: $\bar{x} = \dfrac{198}{12} = 16.5$. The sum of squared deviations is 23.0, so $s = \sqrt{\dfrac{23}{11}} \approx 1.446$. $df = 11$, and from the table at $df = 11$, one-tail area 0.05: $t^* = 1.796$. $SE = \dfrac{1.446}{\sqrt{12}} \approx 0.417$.

$$16.5 \pm 1.796(0.417) = 16.5 \pm 0.75 = (15.75,\ 17.25) \text{ minutes.}$$

Conclude: We are 90% confident that the true mean drying time of this paint formulation is between 15.75 and 17.25 minutes.

VN **GIẢI THÍCH TIẾNG VIỆT**
Công thức khoảng tin cậy t một mẫu luôn có dạng **ước lượng điểm $\pm$ sai số biên**: $\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$. Khi $n < 30$, bắt buộc phải kiểm tra đồ thị (dotplot, boxplot...) để xác nhận không có lệch mạnh hay ngoại lai trước khi áp dụng công thức – đây là bước "Plan" không được bỏ qua, kể cả khi đề bài chỉ cho số liệu tóm tắt.

7.5 The One-Sample t Test for μ

To test a claim $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$ (or μ_0 with a $<$ or $>$), compute

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \qquad df = n - 1,$$

then find (or bound) the P-value using the t-distribution with that df. The reasoning is identical to a one-proportion z-test: assume H_0 is true, measure how many standard errors the observed $\bar{x}$ sits from μ_0, and ask how likely a result this extreme (or more extreme) would be by chance alone.

What a P-value actually means

The P-value is the probability, **assuming H_0 is true**, of obtaining a test statistic at least as extreme (in the direction of H_a) as the one actually observed. It is *not* the probability that H_0 is true, and it is *not* the probability of making an error. A small P-value means the observed data would be surprising if H_0 were true, which is treated as evidence against H_0; it does not by itself measure the size or importance of the effect.

VN **GIẢI THÍCH TIẾNG VIỆT**

P-value là xác suất quan sát được thống kê kiểm định **cực đoan bằng hoặc hơn** giá trị đã tính, **với giả định H_0 đúng**. Đây KHÔNG phải là xác suất H_0 đúng, và cũng KHÔNG phải xác suất mắc sai lầm. P-value nhỏ nghĩa là kết quả quan sát được sẽ rất bất thường nếu H_0 thật sự đúng – đó là bằng chứng chống lại H_0, chứ bản thân P-value không đo độ lớn thực tế của hiệu ứng.

Ví dụ mẫu · A two-sided one-sample t test

A city claims mean commute time is $\mu_0 = 32$ minutes. A random sample of $n = 16$ commuters gives $\bar{x} = 34.5$, $s = 6$ (roughly symmetric, no outliers). Test $H_0 : \mu = 32$ vs. $H_a : \mu \neq 32$ at $\alpha = 0.05$.

State: $H_0 : \mu = 32$ minutes; $H_a : \mu \neq 32$ minutes, where μ is the true mean commute time. $\alpha = 0.05$.

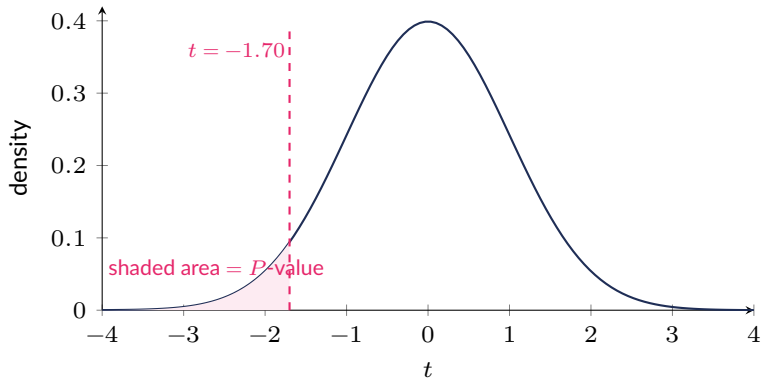
Plan: one-sample t-test, $df = 15$. Random: stated. 10%: 16 commuters is well under 10% of all commuters in the city. Normal/Large Sample: distribution described as roughly symmetric with no outliers, so proceed despite $n < 30$.

Do:

$$t = \frac{34.5 - 32}{6/\sqrt{16}} = \frac{2.5}{1.5} \approx 1.67.$$

From the t-table at $df = 15$: $t = 1.341$ has one-tail area 0.10, and $t = 1.753$ has one-tail area 0.05. Since $1.341 < 1.67 < 1.753$, the one-tail area is between 0.05 and 0.10, so the two-tail P-value is between 0.10 and 0.20.

Conclude: Since $P > 0.10 > \alpha = 0.05$, we **fail to reject** H_0: there is not enough evidence that the true mean commute time differs from 32 minutes.



The sampling distribution of the test statistic (approximately bell-shaped, $df = 19$) under H_0, with the shaded lower tail beyond $t = -1.70$ representing the P-value of a one-sided lower-tail test.

Ví dụ mẫu · A one-sided one-sample t test

An energy-bar manufacturer states that its bars have a mean weight of at least 45 g. A consumer group suspects the bars are underfilled and weighs a random sample of $n = 20$ bars, finding $\bar{x} = 44.2$ g and $s = 2.1$ g (distribution roughly symmetric, no outliers). Test at $\alpha = 0.05$ whether there is evidence the true mean weight is less than 45 g.
State: $H_0 : \mu = 45$ g; $H_a : \mu < 45$ g, where μ is the true mean weight of all bars produced. $\alpha = 0.05$.
Plan: one-sample t-test, $df = 19$. Random: stated. 10%: 20 bars is far less than 10% of total production. Normal/Large Sample: roughly symmetric, no outliers, so proceed despite $n < 30$.
Do:

$$t = \frac{44.2 - 45}{2.1/\sqrt{20}} = \frac{-0.8}{0.4696} \approx -1.70.$$

Since $df = 19$: $t = 1.328$ has one-tail area 0.10, and $t = 1.729$ has one-tail area 0.05. Because $|t| = 1.70$ lies between 1.328 and 1.729, and this is already a one-sided test, the P-value itself is between 0.05 and 0.10.
Conclude: Since $P > 0.05 = \alpha$, we **fail to reject** H_0: there is not enough evidence at the 5% level that the true mean bar weight is below 45 g, although the sample result is suggestive and a larger sample might resolve the question more clearly.

(!) Lỗi thường gặp: Khi đọc bảng t cho một **kiểm định một phía** (one-sided, H_a dùng $<$ hoặc $>$), diện tích đuôi tra được từ bảng **chính là** P-value – KHÔNG được nhân đôi. Ngược lại, với kiểm định **hai phía** (H_a dùng $\neq$), phải nhân đôi diện tích một đuôi để ra P-value hai phía. Nhầm lẫn giữa hai trường hợp này là lỗi rất phổ biến, đặc biệt khi đề bài đổi từ dạng "khác" sang dạng "lớn hơn/nhỏ hơn".

As with any significance test, a decision based on a t-test can be wrong in one of two ways. In the energy-bar example, a **Type I error** would mean rejecting $H_0 : \mu = 45$ (concluding the bars are underfilled) when in fact the true mean weight really is 45 g; a **Type II error** would mean failing to reject H_0 (missing the underfilling) when the true mean weight is actually below 45 g. The significance level α is chosen in advance as the maximum acceptable probability of a Type I error, which is exactly why the decision rule "reject H_0 if $P < \alpha$" is applied consistently rather than adjusted after seeing the data.

VN **GIẢI THÍCH TIẾNG VIỆT**

Thống kê kiểm định $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ đo "khoảng cách" giữa $\bar{x}$ quan sát được và giá trị giả thuyết μ_0, tính theo đơn vị sai số chuẩn. $|t|$ càng lớn thì bằng chứng chống lại H_0 càng mạnh. Vì bảng t chỉ cho một vài mốc diện tích đuôi cố định, ta thường chỉ **chặn khoảng** cho P-value (ví dụ: "P nằm giữa 0.025 và 0.05") thay vì tính chính xác – điều này hoàn toàn đủ để so sánh với α và đưa ra kết luận.

7.5.1 Confidence Intervals and Two-Sided Tests Are Linked

A two-sided significance test and a confidence interval answer closely related questions, and they are mathematically connected: a two-sided test of $H_0 : \mu = \mu_0$ at significance level α rejects H_0 if and only if μ_0 falls *outside* the corresponding $(1 - \alpha)$ confidence interval for μ, computed from the same data.

> For a two-sided test at significance level α: reject $H_0 : \mu = \mu_0$ if and only if μ_0 lies outside the $(1 - \alpha)$ confidence interval for μ. Equivalently, fail to reject H_0 if and only if μ_0 lies inside that interval.

Ví dụ mẫu · Using a CI to answer a two-sided test

Using the light-bulb data from earlier ($n = 25$, $\bar{x} = 850$, $s = 40$), the 95% CI for μ was $(833.5,\ 866.5)$ hours. Without recomputing a t-statistic, test $H_0 : \mu = 840$ vs. $H_a : \mu \neq 840$ at $\alpha = 0.05$.

Since 840 lies inside the interval $(833.5,\ 866.5)$, a two-sided test at $\alpha = 0.05$ would **fail to reject** $H_0 : \mu = 840$: the 95% CI is entirely consistent with a true mean lifetime of 840 hours. (Had the claim instead been $\mu_0 = 870$, which lies outside $(833.5,\ 866.5)$, the same reasoning would reject H_0 at $\alpha = 0.05$.)

VN **GIẢI THÍCH TIẾNG VIỆT**

Có mối liên hệ hai chiều giữa **khoảng tin cậy** và **kiểm định hai phía**: nếu μ_0 nằm NGOÀI khoảng tin cậy mức $(1-\alpha)$, kiểm định hai phía ở mức α sẽ bác bỏ H_0; nếu μ_0 nằm TRONG khoảng, kiểm định sẽ không bác bỏ H_0. Mẹo này giúp trả lời nhanh nhiều câu hỏi trắc nghiệm mà không cần tính lại thống kê t từ đầu – miễn là mức ý nghĩa α và mức tin cậy $(1-\alpha)$ khớp nhau.

7.6 Two-Sample t Procedures for $\mu_1 - \mu_2$

Two independent samples: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; a conservative df is $\min(n_1 - 1, n_2 - 1)$. The same standard error, $SE = \sqrt{s_1^2/n_1 + s_2^2/n_2}$, is used both for a confidence interval and for a significance test.

Some textbooks and software also offer a **pooled two-sample t procedure**, which assumes the two populations share a common variance $\sigma_1^2 = \sigma_2^2 = \sigma_p^2$ and combines s_1 and s_2 into a single pooled estimate with $df = n_1 + n_2 - 2$. AP Statistics does not use the pooled procedure: the equal-variance assumption is difficult to verify and rarely worth the risk, while the *unpooled* approach used throughout this unit performs nearly as well when variances happen to be equal and considerably better when they are not. Always use the unpooled formula $SE = \sqrt{s_1^2/n_1 + s_2^2/n_2}$ with conservative $df = \min(n_1 - 1, n_2 - 1)$ for hand calculations in this course.

7.6.1 Confidence Interval for $\mu_1 - \mu_2$

A two-sample t interval estimates the true difference in population means, $\mu_1 - \mu_2$, using $(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot SE$. Conditions require randomness (or random assignment) and independence *both within and between* the two samples – this is what distinguishes the procedure from matched pairs. Whether the "Random" condition is satisfied by random *sampling* (an observational study, generalizing only to the sampled population) or by random *assignment* to treatments (a randomized experiment, supporting a cause-and-effect conclusion) changes what kind of conclusion the final answer is allowed to make, even though the arithmetic of the interval itself is identical either way.

Ví dụ mẫu · A two-sample t confidence interval

Two teaching methods are compared using independent classes. Method A: $n_1 = 10$, $\bar{x}_1 = 82$, $s_1 = 6$. Method B: $n_2 = 12$, $\bar{x}_2 = 77$, $s_2 = 8$. Construct a 95% confidence interval for $\mu_A - \mu_B$.

State: parameter is $\mu_A - \mu_B$, the true difference in mean scores between students taught by Method A and by Method B. Target: 95% CI.

Plan: two-sample t interval with conservative $df = \min(9, 11) = 9$. Random: assume classes represent random/independently assigned groups. 10%: both samples are small relative to the population of all students who could take these courses. Normal/Large Sample: assume

the score distributions are roughly symmetric with no outliers (as is typical for exam scores of this kind), so we may proceed despite $n_1, n_2 < 30$. The two groups are independent of one another.
Do: $t^* = 2.262$ ($df = 9$, one-tail area 0.025).

$$SE = \sqrt{\frac{6^2}{10} + \frac{8^2}{12}} = \sqrt{3.6 + 5.333} = \sqrt{8.933} \approx 2.989.$$

$$ME = 2.262(2.989) \approx 6.76. \qquad (82 - 77) \pm 6.76 = 5 \pm 6.76 = (-1.76,\ 11.76).$$

Conclude: We are 95% confident that the true difference in mean scores $\mu_A - \mu_B$ is between -1.76 and 11.76 points. Because this interval contains 0, it is plausible that the two methods produce the same population mean score – the data do not provide convincing evidence of a difference.

7.6.2 Significance Test for $\mu_1 - \mu_2$

A two-sample t-test checks $H_0 : \mu_1 - \mu_2 = 0$ (equivalently $\mu_1 = \mu_2$) against a one- or two-sided alternative, using the same standardized statistic $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{SE}$ and the same conservative df.

Ví dụ mẫu · A two-sample t test

Two independent random samples compare the mean battery life (hours) of two smartphone models. Model X: $n_1 = 14$, $\bar{x}_1 = 11.8$, $s_1 = 1.5$. Model Y: $n_2 = 11$, $\bar{x}_2 = 10.6$, $s_2 = 1.9$. Both distributions are roughly symmetric with no outliers. Test at $\alpha = 0.05$ whether there is a significant difference in mean battery life.
State: $H_0 : \mu_X - \mu_Y = 0$; $H_a : \mu_X - \mu_Y \neq 0$, where μ_X, μ_Y are the true mean battery lives of the two models. $\alpha = 0.05$.
Plan: two-sample t-test, conservative $df = \min(13, 10) = 10$. Random: independent random samples stated. 10%: satisfied for both models. Normal/Large Sample: both distributions roughly symmetric with no outliers, so proceed despite $n < 30$. Samples are independent of each other.
Do:

$$SE = \sqrt{\frac{1.5^2}{14} + \frac{1.9^2}{11}} = \sqrt{0.1607 + 0.3282} = \sqrt{0.4889} \approx 0.6992.$$

$$t = \frac{11.8 - 10.6}{0.6992} = \frac{1.2}{0.6992} \approx 1.72.$$

At $df = 10$: $t = 1.372$ has one-tail area 0.10; $t = 1.812$ has one-tail area 0.05. Since $1.372 < 1.72 < 1.812$, the one-tail area is between 0.05 and 0.10, so the two-tail P-value is between 0.10 and 0.20.
Conclude: Since $P > 0.10 > \alpha = 0.05$, we **fail to reject** H_0: there is not enough evidence of a significant difference in mean battery life between the two models.

VN **GIẢI THÍCH TIẾNG VIỆT**
Thủ tục **hai mẫu độc lập** dùng khi hai nhóm dữ liệu đến từ những đối tượng **khác nhau, không liên quan**, được lấy mẫu (hoặc phân nhóm) độc lập với nhau. Công thức $SE = \sqrt{s_1^2/n_1 + s_2^2/n_2}$ cộng phương sai (không phải độ lệch chuẩn) của hai mẫu – một lỗi thường gặp là cộng trực tiếp $s_1 + s_2$ hoặc $s_1/n_1 + s_2/n_2$ mà quên bình phương và căn bậc hai đúng thứ tự.

(!) Lỗi thường gặp: Với hai mẫu độc lập, luôn dùng df **bảo thủ** $\min(n_1 - 1, n_2 - 1)$ khi tính tay (phần mềm dùng công thức Welch chính xác hơn, cho df lẻ và thường lớn hơn) – không được cộng $n_1 + n_2 - 2$ như hồi quy hay ANOVA. Dùng df bảo thủ luôn cho kết quả **an toàn hơn** (khoảng tin cậy rộng hơn một chút, P-value lớn hơn một chút) so với giá trị df chính xác của Welch.

7.6.3 Reading Statistical Software Output

AP Statistics exam questions and real research reports often present computer output rather than raw data. Locating the test statistic, degrees of freedom, and P-value inside a printout – and reconciling them with a hand calculation – is its own essential skill.

Ví dụ mẫu · Interpreting two-sample t software output

A researcher compares reaction time (ms) under Caffeine vs. Placebo using two independent groups of 15 participants each and reports the following software output:

Group	N	Mean	StDev	SE Mean
Caffeine	15	342.6	28.4	7.3
Placebo	15	361.2	31.7	8.2

Difference $= -18.6$ T-Value $= -1.69$ DF $= 27$ P-Value $= 0.102$

Use this output to test $H_0 : \mu_C - \mu_P = 0$ vs. $H_a : \mu_C - \mu_P \neq 0$ at $\alpha = 0.05$.

The software uses the more precise Welch approximation, giving a non-conservative $DF = 27$ rather than the hand-calculation value $df = \min(14, 14) = 14$; both approaches lead to the same conclusion here. Since $P = 0.102 > \alpha = 0.05$, we **fail to reject** H_0: this sample does not provide convincing evidence that mean reaction time differs between Caffeine and Placebo. *Checking by hand* with the conservative $df = 14$: $t = -1.69$ falls between the $df = 14$ table values 1.345 (one-tail area 0.10) and 1.761 (one-tail area 0.05), so the two-tail P-value is bounded between 0.10 and 0.20 – consistent with, though less precise than, the software's $P = 0.102$.

VN **GIẢI THÍCH TIẾNG VIỆT**

Khi đọc kết quả xuất ra từ phần mềm thống kê, cần xác định đúng bốn thành phần: cỡ mẫu và thống kê mô tả của từng nhóm, giá trị kiểm định (T-Value), bậc tự do (DF – thường là số lẻ do dùng công thức Welch), và P-value. Luôn có thể **đối chiếu ngược** bằng cách tính tay với df bảo thủ để kiểm tra tính hợp lý của kết quả phần mềm, như trong ví dụ trên.

Statistical significance answers only one narrow question: is the observed difference unlikely to have arisen from chance alone? It says nothing about whether that difference is large enough to matter in practice. With a very large sample, even a tiny, practically meaningless difference can become statistically significant (a very small P-value); with a small sample, a genuinely important difference can fail to reach significance. For this reason, a complete answer always reports and interprets the point estimate and confidence interval alongside the P-value, so that practical significance can be judged separately from statistical significance.

VN **GIẢI THÍCH TIẾNG VIỆT**

Ý nghĩa thống kê (statistical significance) chỉ trả lời câu hỏi "khác biệt quan sát được có khó xảy ra do ngẫu nhiên hay không?" – KHÔNG nói lên khác biệt đó có **đủ lớn để có ý nghĩa thực tế** (practical significance) hay không. Mẫu càng lớn thì càng dễ phát hiện những khác biệt rất

nhỏ là "có ý nghĩa thống kê", dù khác biệt đó có thể chẳng quan trọng gì trong thực tế. Vì vậy, luôn báo cáo kèm khoảng tin cậy và độ lớn khác biệt thực tế, không chỉ riêng P-value.

7.7 Matched Pairs t Procedures

7.7.1 Reducing Pairs to a One-Sample Problem

Matched pairs: first compute the differences d_i for each pair, then run a *one-sample* t procedure on the differences with $df = n_{\text{pairs}} - 1$. All formulas from Sections 4–5 apply directly, with $\bar{x}$ replaced by $\bar{d}$, s replaced by s_d, and μ replaced by μ_d (the true mean difference).

A common point of confusion: because a matched-pairs design produces $2n$ individual measurements (before and after, or two conditions per pair), it is tempting to use $df = 2n - 2$ as though this were a two-sample procedure. This is incorrect – once the data are reduced to the n differences d_i, there are only n data points left to analyze, so $df = n - 1$, never $2n - 2$ or $2n - 1$. The Random condition for matched pairs can be satisfied either by randomly sampling the pairs themselves, or, in an experiment, by randomly assigning which member of each pair receives which treatment (as with a study that randomly assigns one twin in each pair to a treatment and the other to a control) – the pairing itself need not be random, only the sampling or assignment.

VN **GIẢI THÍCH TIẾNG VIỆT**

Bài toán **matched pairs** (bắt cặp) luôn quy về kiểm định/khoảng tin cậy **một mẫu** trên các **hiệu số** d_i – không dùng công thức hai mẫu độc lập. Dấu hiệu nhận biết: "trước–sau" trên cùng một đối tượng, hoặc từng cặp được ghép chủ đích (ví dụ hai giày trái/phải của cùng một người, hoặc hai anh em sinh đôi).

Pairing is a deliberate design choice, and it usually pays off. When the two measurements within a pair are positively correlated – a subject who starts relatively heavy tends to remain relatively heavy; a naturally fast runner tends to stay relatively fast – computing the difference d_i for each pair cancels out much of that shared, subject-specific variability before it ever reaches s_d. The result is often a considerably smaller standard error than an independent two-sample design with the same total number of measurements would produce, which is exactly why researchers deliberately pair or block subjects whenever a natural pairing is available.

Ví dụ mẫu · A matched-pairs t test

Eight participants' weights (lb) before/after an 8-week program: differences (before − after) are $6, 3, 7, 4, 3, 9, 6, 2$. Test whether the program reduces mean weight: $H_0 : \mu_d = 0$ vs. $H_a : \mu_d > 0$, at $\alpha = 0.01$.

State: $H_0 : \mu_d = 0$; $H_a : \mu_d > 0$, where μ_d is the true mean difference (before − after) in weight for participants of this program. $\alpha = 0.01$.

Plan: matched-pairs t-test, $df = n_{\text{pairs}} - 1 = 7$. Random: assume participants are a random sample of the target population. 10%: satisfied. Normal/Large Sample: with $n = 8 < 15$, we need the differences to show no strong skew or outliers; assume this is confirmed by a dotplot of the differences.

Do: $\bar{d} = \dfrac{6+3+7+4+3+9+6+2}{8} = \dfrac{40}{8} = 5$. Deviations: $1, -2, 2, -1, -2, 4, 1, -3$; squares sum to 40; $s_d = \sqrt{40/7} \approx 2.39$.

$$t = \frac{5 - 0}{2.39/\sqrt{8}} = \frac{5}{0.845} \approx 5.92, \qquad df = 7.$$

The t-table value for $df = 7$ at one-tail area 0.005 is 3.499; since $5.92 > 3.499$, $P < 0.005$.

Conclude: Since $P < 0.005 < \alpha = 0.01$, we **reject** H_0: there is strong evidence that the program reduces mean weight.

(!) Lỗi thường gặp: Lỗi thường gặp nhất với matched pairs: học sinh dùng **công thức hai mẫu độc lập** $SE = \sqrt{s_1^2/n_1 + s_2^2/n_2}$ trên dữ liệu "trước" và "sau" thay vì tính hiệu số d_i trước rồi áp dụng công thức một mẫu. Cách làm sai này bỏ qua sự **tương quan dương** giữa "trước" và "sau" trên cùng một đối tượng (người có cân nặng ban đầu cao thường vẫn cao sau chương trình), khiến SE bị tính **quá lớn** một cách không cần thiết và làm giảm sức mạnh của kiểm định.

7.7.2 Confidence Interval for the Mean Difference μ_d

The same reduction to differences applies to confidence intervals: compute $\bar{d} \pm t^* \cdot \frac{s_d}{\sqrt{n}}$ with $df = n - 1$, where n is the number of pairs.

Ví dụ mẫu · A matched-pairs t confidence interval

A computer-skills instructor records typing speed (words per minute) for $n = 10$ students before and after a 4-week typing course. The differences (after − before) are:

$$5,\ 3,\ 6,\ 4,\ 2,\ 7,\ 5,\ 4,\ 6,\ 3.$$

Construct a 95% confidence interval for the true mean improvement μ_d.
State: parameter is μ_d, the true mean improvement (after − before) in typing speed for students who take this course. Target: 95% CI.
Plan: matched-pairs t interval, $df = n - 1 = 9$. Random: students assumed to be a random sample of the target population. 10%: satisfied. Normal/Large Sample: $n = 10 < 30$; assume a dotplot of the differences shows no strong skew or outliers.
Do: $\bar{d} = \frac{45}{10} = 4.5$. Sum of squared deviations from $\bar{d}$ is 22.5, so $s_d^2 = \frac{22.5}{9} = 2.5$ and $s_d = \sqrt{2.5} \approx 1.581$. $t^* = 2.262$ ($df = 9$, one-tail area 0.025). $SE = \frac{1.581}{\sqrt{10}} = 0.500$.

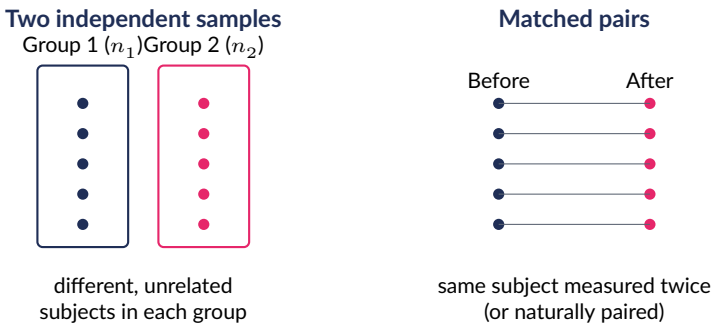
$$4.5 \pm 2.262(0.500) = 4.5 \pm 1.131 = (3.37,\ 5.63) \text{ wpm}.$$

Conclude: We are 95% confident that the true mean improvement in typing speed after this course is between 3.37 and 5.63 words per minute. Because the entire interval lies above 0, the data provide convincing evidence of a genuine (positive) improvement.

(!) Lỗi thường gặp: Trước khi tính d_i, phải chọn và **giữ nhất quán** một quy ước dấu – ví dụ "sau − trước" hoặc "trước − sau" – rồi áp dụng *đúng* quy ước đó cho cả $\bar{d}$ lẫn hướng của H_a. Đổi chiều quy ước giữa lúc tính $\bar{d}$ và lúc viết H_a (ví dụ dùng "trước − sau" để tính số liệu nhưng lại viết $H_a : \mu_d > 0$ với ý nghĩa "sau lớn hơn trước") sẽ khiến dấu của t và kết luận cuối cùng bị đảo ngược so với thực tế.

7.7.3 Two-Sample or Matched Pairs? Choosing the Right Procedure

The single most important decision before running any two-group t procedure is recognizing *which* procedure applies. The test is structural, not statistical: ask whether the two sets of measurements are linked observation-by-observation.



Independent samples (left) compare separate, unrelated groups of subjects; matched pairs (right) link each measurement in one set to a specific, corresponding measurement in the other set.

Ví dụ mẫu · Two-sample or matched pairs? Three scenarios

Scenario 1. A researcher randomly assigns 40 volunteers into two independent groups: 20 take Supplement A, 20 take Supplement B, and cholesterol levels are compared after 8 weeks.
→ **Two independent samples.** The volunteers in Group A share no special link with those in Group B; each group is a separate, unrelated set of subjects.

Scenario 2. A physical therapist measures the grip strength of 15 patients before and after a 6-week rehabilitation program.
→ **Matched pairs.** Each patient contributes two linked measurements (before and after), so the natural unit of analysis is the within-patient difference d_i.

Scenario 3. A shoe company tests the durability of two sole materials by having each of 12 runners wear one shoe of each material (one on each foot, randomly assigned to left or right) for 500 miles.
→ **Matched pairs.** Even though two different materials are compared (not the same subject twice), each runner acts as a block that pairs one measurement of Material 1 with one measurement of Material 2 – the same logic as "before/after" pairing.

VN **GIẢI THÍCH TIẾNG VIỆT**

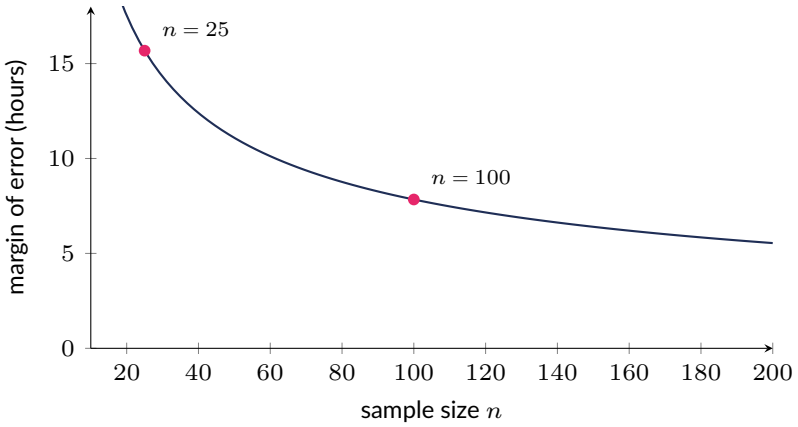
Câu hỏi quyết định: **"Mỗi giá trị ở nhóm 1 có được ghép tự nhiên với đúng một giá trị ở nhóm 2 hay không?"** Nếu CÓ (cùng một đối tượng đo hai lần, hoặc hai đối tượng được ghép chủ đích như cặp song sinh, hai chân của cùng một người) ⇒ **matched pairs**. Nếu KHÔNG (hai nhóm gồm những đối tượng hoàn toàn tách biệt, được chọn/phân ngẫu nhiên độc lập) ⇒ **hai mẫu độc lập**. Chọn sai thủ tục sẽ cho SE sai và toàn bộ phần "Do" bị sai theo.

Putting the three procedures side by side. Every t procedure in this unit follows the same two patterns – CI: statistic $\pm t^* \cdot SE$; test: $t = \dfrac{\text{estimate} - \text{hypothesized value}}{SE}$ – with the parameter, standard error, and degrees of freedom changing as shown below.

Procedure	Parameter	Standard error	Degrees of freedom
One-sample	μ	$\dfrac{s}{\sqrt{n}}$	$n-1$
Two-sample (independent)	$\mu_1 - \mu_2$	$\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$	$\min(n_1-1, n_2-1)$ (conservative)
Matched pairs	μ_d	$\dfrac{s_d}{\sqrt{n_{\text{pairs}}}}$	$n_{\text{pairs}} - 1$

7.8 Sample Size and the Margin of Error

The margin of error of a t interval, $ME = t^* \cdot \dfrac{s}{\sqrt{n}}$, shrinks as n grows – but not proportionally to n. Because n appears inside a square root, ME is proportional to $\dfrac{1}{\sqrt{n}}$: to cut the margin of error in half, the sample size must be roughly **quadrupled**, not merely doubled. (The critical value t^* also decreases slightly as n grows, making the actual reduction in ME marginally better than the $1/\sqrt{n}$ rule alone would suggest.)



Margin of error versus sample size, using $s = 40$ and $z^* \approx 1.96$ as an approximation for t^* at large df. Quadrupling n from 25 to 100 roughly halves the margin of error.

Ví dụ mẫu · Margin of error shrinks with n

Reconsider the light-bulb scenario with $s = 40$ hours. With $n = 25$, the 95% margin of error was $t^* \cdot SE = 2.064(8) \approx 16.51$ hours. Suppose instead $n = 100$. For $df = 99$, the critical value t^* is very close to $z^* = 1.960$ (the $df = \infty$ row), so

$$SE = \frac{40}{\sqrt{100}} = 4, \qquad ME \approx 1.960(4) = 7.84 \text{ hours.}$$

Quadrupling the sample size ($25 \to 100$) exactly halves the standard error ($8 \to 4$, since $\sqrt{100}/\sqrt{25} = 2$), and the margin of error drops from about 16.51 to about 7.84 hours – slightly more than a straight halving, because t^* also fell slightly (from 2.064 toward 1.960) as df increased.

Ví dụ mẫu · Planning a sample size

Pilot data suggest $s \approx 40$ hours (light-bulb context). Researchers want a 95% confidence interval with margin of error no more than 10 hours. Using $z^* \approx 1.96$ as a first approximation (the exact df is not yet known, since it depends on the sample size being solved for), solve $ME = z^* \cdot \dfrac{s}{\sqrt{n}} \le 10$ for n:

$$n \ge \left(\frac{z^* s}{ME}\right)^2 = \left(\frac{1.96(40)}{10}\right)^2 = (7.84)^2 = 61.47.$$

Since n must be a whole number and larger n only helps, round **up**: $n = 62$. (With $n = 62$, $df = 61$ is close to the $df = 40$ row of the table, where $t^* = 2.021$ – still near 1.96, confirming the approximation was reasonable.)

In practice, the value of s used for planning almost never comes from thin air – it is typically borrowed from a small pilot study, a similar study published elsewhere, or a known range for the variable being measured (for a roughly Normal variable, a rough range-based estimate is $s \approx \text{range}/4$). Because required sample size grows with the *square* of both z^* and s, but only shrinks with the square of ME, tightening precision or raising confidence is expensive: doubling the confidence level's z^* or doubling s quadruples the required n, while halving the target margin of error also quadruples it. Researchers therefore treat sample-size planning as a genuine trade-off between the cost of collecting more data and the value of a tighter interval, not a step to maximize blindly.

VN **GIẢI THÍCH TIẾNG VIỆT**

Vì $ME = t^* \cdot s/\sqrt{n}$, muốn giảm sai số biên xuống một nửa thì phải tăng cỡ mẫu lên **gấp bốn lần** (không phải gấp đôi) – do n nằm trong căn bậc hai. Khi lập kế hoạch chọn cỡ mẫu trước khi thu thập dữ liệu, người ta thường dùng z^* (thay vì t^*) làm xấp xỉ ban đầu, vì df thật sự phụ thuộc vào chính n đang cần tìm – một bài toán "con gà và quả trứng" mà xấp xỉ bằng z^* giải quyết gọn gàng. Trong thực tế, s dùng để lập kế hoạch thường lấy từ một nghiên cứu thử nghiệm nhỏ (pilot study) hoặc từ tài liệu tương tự đã công bố; vì cỡ mẫu cần thiết tăng theo **bình phương** của cả z^* lẫn s, việc đòi hỏi độ tin cậy cao hơn hoặc sai số biên nhỏ hơn luôn phải đánh đổi bằng chi phí thu thập dữ liệu lớn hơn đáng kể.

(!) Lỗi thường gặp: Trên đề tự luận (FRQ), điểm bị mất nhiều nhất thường đến từ việc **bỏ qua bước "Plan"**: viết đúng công thức và ra đúng số ở bước "Do" KHÔNG tự động được điểm cho phần kiểm tra điều kiện. Mọi bài toán một mẫu, hai mẫu, hay matched pairs đều phải nêu rõ **Random**, **10%**, và **Normal/Large Sample** (hoặc điều kiện tương ứng), đồng thời giải thích từng điều kiện dựa trên dữ kiện đề bài cho – không chỉ khẳng định suông rằng điều kiện "đã thoả mãn".

7.9 Practice Problems

Bài tập luyện tập

Two teaching methods are compared. Method A: $n_1 = 10$, $\bar{x}_1 = 82$, $s_1 = 6$. Method B: $n_2 = 12$, $\bar{x}_2 = 77$, $s_2 = 8$. Construct a 95% CI for $\mu_A - \mu_B$ using conservative $df = 9$ ($t^* = 2.262$).

(A) $5 \pm 6.76 = (-1.76,\ 11.76)$

(B) $5 \pm 2.26 = (2.74,\ 7.26)$

(C) $-5 \pm 6.76 = (-11.76,\ 1.76)$

(D) $5 \pm 1.41 = (3.59,\ 6.41)$

Lời giải chi tiết

$SE = \sqrt{\frac{6^2}{10} + \frac{8^2}{12}} = \sqrt{3.6 + 5.333} = \sqrt{8.933} \approx 2.989$. $ME = 2.262(2.989) \approx 6.76$. CI: $(82 - 77) \pm 6.76 = 5 \pm 6.76 = (-1.76, 11.76)$. **(A)**. Since the interval contains 0, we cannot conclude a significant difference in mean scores at the 95% level.

Bài tập luyện tập

Which is the correct condition to check before a one-sample t-test with $n = 12$ and a dotplot showing one clear high outlier?

(A) Large Counts condition, since $n < 30$

(B) Normal/Large Sample condition may fail – the outlier and small n make the t-procedure unreliable

(C) No condition is needed since t-procedures are always robust

(D) The 10% condition, since the outlier affects independence

Lời giải chi tiết

With small n and a clear outlier, the sampling distribution of $\bar{x}$ may not be close to Normal, and t-procedures are not robust to strong outliers at small sample sizes. **(B)**.

Bài tập luyện tập

A 90% confidence interval for mean battery life is $(38.2,\ 41.8)$ hours. Which is a correct interpretation?

(A) 90% of batteries last between 38.2 and 41.8 hours

(B) There is a 90% probability the true mean is in $(38.2, 41.8)$

(C) We used a method that captures the true mean in about 90% of all possible samples; we are 90% confident this particular interval captures μ

(D) The sample mean is 90% likely to be 40

Lời giải chi tiết

A confidence level describes the long-run capture rate of the *method* across repeated sampling, translated to "confidence" for the one interval computed. **(C)**. Options (A) and (D) confuse individual values with the mean; (B) misapplies probability to a fixed parameter.

Bài tập luyện tập

For a 99% confidence interval based on a sample of $n = 11$, what are df and t^*?

(A) $df = 10, t^* = 3.169$

(B) $df = 11, t^* = 3.106$

(C) $df = 10, t^* = 2.764$

(D) $df = 10, t^* = 2.228$

Lời giải chi tiết

$df = n - 1 = 10$. A 99% CI leaves 0.01 total in the tails, so each tail has area 0.005. At $df = 10$, the 0.005 column gives $t^* = 3.169$. **(A)**. (B) uses the wrong df; (C) and (D) use the wrong column (0.01 and 0.025).

Bài tập luyện tập

Compared to the standard Normal distribution, the t-distribution with small df is:

(A) more spread out with heavier tails, converging to Normal as df increases

(B) narrower with lighter tails than the Normal

(C) identical in shape for every df

(D) skewed to the right

Lời giải chi tiết

Substituting s for the unknown σ adds extra variability, so the t-distribution has a lower peak and heavier tails than z, especially for small df; as $df \to \infty$, it converges to the standard Normal curve. **(A)**.

Bài tập luyện tập

A sample of $n = 8$ measurements shows a strong left skew with no outliers. Can a one-sample t-test be used safely? Explain.

Lời giải chi tiết

Not safely without further caution. With $n = 8$ (well under 15) and clear strong skew, the Normal/Large Sample condition is doubtful: t-procedures are not robust to strong skewness at such small sample sizes, since the sampling distribution of $\bar{x}$ may itself be noticeably skewed rather than approximately Normal. A larger sample, a transformation, or a nonparametric procedure would be more appropriate.

Bài tập luyện tập

A sample of $n = 50$ is strongly right-skewed with no extreme outliers. Is the Normal/Large Sample condition met? Explain.

Lời giải chi tiết

Yes. Because $n = 50 \geq 30$, the Central Limit Theorem guarantees that the sampling distribution of $\bar{x}$ is approximately Normal regardless of the skew present in the population, as long as there are no extreme outliers distorting $\bar{x}$ and s.

Bài tập luyện tập

Which scenario best satisfies the Normal/Large Sample condition for a one-sample t-test?

(A) $n = 10$ with one clear outlier

(B) $n = 15$, roughly symmetric, no outliers

(C) $n = 5$, clearly and strongly skewed

(D) $n = 8$, strongly bimodal

Lời giải chi tiết

With $n < 30$, we need the sample data to show no strong skew and no outliers. Only the $n = 15$, roughly symmetric, no-outlier sample satisfies this. **(B)**.

Bài tập luyện tập

A random sample of $n = 20$ measures the caffeine content (mg) of a new energy drink: $\bar{x} = 95.4$ mg, $s = 8.2$ mg (roughly symmetric, no outliers). Construct a 90% confidence interval for the true mean caffeine content μ.

Lời giải chi tiết

$df = 19$. For 90% confidence, one-tail area $= 0.05$, so $t^* = 1.729$. $SE = \dfrac{8.2}{\sqrt{20}} \approx 1.834$.

$ME = 1.729(1.834) \approx 3.17$. CI: $95.4 \pm 3.17 = (92.23,\ 98.57)$ mg. We are 90% confident the true mean caffeine content is between 92.23 and 98.57 mg.

Bài tập luyện tập

A cereal box states the mean weight is 500 g. A random sample of $n = 30$ boxes gives $\bar{x} = 496$ g, $s = 12$ g. Test $H_0 : \mu = 500$ vs. $H_a : \mu \neq 500$ at $\alpha = 0.05$.

Lời giải chi tiết

$df = 29$. $t = \dfrac{496 - 500}{12/\sqrt{30}} = \dfrac{-4}{2.191} \approx -1.83$. At $df = 29$: $t = 1.699$ has one-tail area 0.05; $t = 2.045$ has one-tail area 0.025. Since $1.699 < 1.83 < 2.045$, the one-tail area is between 0.025 and 0.05, so the two-tail P-value is between 0.05 and 0.10. Since $P > 0.05 = \alpha$, we **fail to reject** H_0: not enough evidence the true mean box weight differs from 500 g (a close call).

Bài tập luyện tập

A tire manufacturer claims its new tire lasts more than $50{,}000$ miles on average. A random sample of $n = 16$ tires gives $\bar{x} = 51{,}200$ miles, $s = 2{,}400$ miles (roughly symmetric, no outliers). Test the claim at $\alpha = 0.05$.

Lời giải chi tiết

State: $H_0 : \mu = 50{,}000$; $H_a : \mu > 50{,}000$ miles. $\alpha = 0.05$.
Plan: one-sample t-test, $df = 15$. Random, 10%, and Normal/Large Sample (roughly symmetric, no outliers, $n < 30$ but condition satisfied) are all met.
Do: $t = \dfrac{51200 - 50000}{2400/\sqrt{16}} = \dfrac{1200}{600} = 2.00$. At $df = 15$: $t = 1.753$ has one-tail area 0.05; $t = 2.131$ has one-tail area 0.025. Since $1.753 < 2.00 < 2.131$, P is between 0.025 and 0.05.
Conclude: Since $P < 0.05 = \alpha$, we **reject** H_0: there is convincing evidence that the true mean tire life exceeds $50{,}000$ miles.

Bài tập luyện tập

A dietitian records the sugar content (g) of $n = 9$ randomly selected granola bars from a new brand: $12, 14, 13, 15, 11, 14, 13, 12, 16$. A dotplot shows a roughly symmetric, single-peaked shape with no outliers. Construct a 95% CI for the true mean sugar content μ.

Lời giải chi tiết

$\bar{x} = \dfrac{120}{9} \approx 13.33$ g. Sum of squared deviations is 20.0, so $s = \sqrt{20/8} \approx 1.581$ g. $df = 8$, $t^* = 2.306$ (one-tail area 0.025). $SE = \dfrac{1.581}{3} \approx 0.527$. $ME = 2.306(0.527) \approx 1.22$. CI: $13.33 \pm 1.22 = (12.11,\ 14.55)$ g. We are 95% confident the true mean sugar content is between 12.11 and 14.55 g.

Bài tập luyện tập

Independent samples: $n_1 = 8$, $\bar{x}_1 = 64$, $s_1 = 5$; $n_2 = 10$, $\bar{x}_2 = 59$, $s_2 = 7$. Construct a 90% CI for $\mu_1 - \mu_2$ using the conservative df.

(A) $5 \pm 5.37 = (-0.37,\ 10.37)$
(C) $-5 \pm 5.37 = (-10.37,\ 0.37)$
(B) $5 \pm 2.32 = (2.68,\ 7.32)$
(D) $5 \pm 8.99 = (-3.99,\ 13.99)$

Lời giải chi tiết

$df = \min(7, 9) = 7$, so $t^* = 1.895$ (one-tail area 0.05). $SE = \sqrt{\frac{25}{8} + \frac{49}{10}} = \sqrt{8.025} \approx 2.833$. $ME = 1.895(2.833) \approx 5.37$. CI: $(64 - 59) \pm 5.37 = 5 \pm 5.37 = (-0.37, 10.37)$. **(A)**.

Bài tập luyện tập

A gym claims a new workout program (Group 1) produces greater average fat loss than a standard program (Group 2). Independent samples: Group 1: $n_1 = 12$, $\bar{x}_1 = 8.4$ kg, $s_1 = 2.1$ kg. Group 2: $n_2 = 10$, $\bar{x}_2 = 6.9$ kg, $s_2 = 1.8$ kg. Test at $\alpha = 0.05$ whether Group 1's mean fat loss is greater.

Lời giải chi tiết

State: $H_0 : \mu_1 - \mu_2 = 0$; $H_a : \mu_1 - \mu_2 > 0$. $\alpha = 0.05$.
Plan: two-sample t-test, conservative $df = \min(11, 9) = 9$. Random, 10%, Normal/Large Sample, and independence between groups are all assumed satisfied.
Do: $SE = \sqrt{\frac{2.1^2}{12} + \frac{1.8^2}{10}} = \sqrt{0.3675 + 0.324} = \sqrt{0.6915} \approx 0.8316$. $t = \frac{8.4 - 6.9}{0.8316} \approx 1.80$. At $df = 9$: $t = 1.383$ has one-tail area 0.10; $t = 1.833$ has one-tail area 0.05. Since $1.383 < 1.80 < 1.833$, P is between 0.05 and 0.10.
Conclude: Since $P > 0.05 = \alpha$, we **fail to reject** H_0: not quite enough evidence that Group 1's mean fat loss exceeds Group 2's, though the result is suggestive.

Bài tập luyện tập

When calculating a two-sample t procedure by hand using the conservative method, the degrees of freedom equal:

(A) $n_1 + n_2 - 2$
(C) $n_1 + n_2 - 1$
(B) $\min(n_1 - 1,\ n_2 - 1)$
(D) $\max(n_1, n_2) - 1$

Lời giải chi tiết

The conservative (by-hand) approach for two independent samples uses $df = \min(n_1 - 1, n_2 - 1)$; the pooled formula $n_1 + n_2 - 2$ belongs to different procedures (e.g., pooled-variance methods, regression, ANOVA) that assume equal population variances. **(B)**.

Bài tập luyện tập

Independent random samples compare mean study hours per week for two dorms. Dorm A: $n_1 = 16$, $\bar{x}_1 = 14.2$, $s_1 = 3.5$. Dorm B: $n_2 = 20$, $\bar{x}_2 = 11.8$, $s_2 = 4.1$ (both roughly symmetric, no outliers). Construct a 95% CI for $\mu_A - \mu_B$.

Lời giải chi tiết

State: parameter $\mu_A - \mu_B$; target 95% CI.
Plan: two-sample t interval, conservative $df = \min(15, 19) = 15$, so $t^* = 2.131$. Random, 10%, Normal/Large Sample, and independence between dorms assumed satisfied.
Do: $SE = \sqrt{\frac{3.5^2}{16} + \frac{4.1^2}{20}} = \sqrt{0.7656 + 0.8405} = \sqrt{1.6061} \approx 1.267$. $ME = 2.131(1.267) \approx 2.70$. CI: $(14.2 - 11.8) \pm 2.70 = 2.4 \pm 2.70 = (-0.30,\ 5.10)$ hours.
Conclude: We are 95% confident the true difference in mean weekly study hours is between -0.30 and 5.10 hours; because the interval contains 0, we cannot conclude a significant difference at this level.

Bài tập luyện tập

Two independent samples of exam scores: $n_1 = 9$, $\bar{x}_1 = 78$, $s_1 = 6$; $n_2 = 9$, $\bar{x}_2 = 83$, $s_2 = 5.5$. Test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$ using conservative $df = 8$. Which is closest to the test statistic and conclusion at $\alpha = 0.05$?

(A) $t \approx -1.84$, $P > 0.05$, fail to reject H_0
(B) $t \approx -1.84$, $P < 0.05$, reject H_0
(C) $t \approx -2.50$, $P < 0.05$, reject H_0
(D) $t \approx -0.91$, $P > 0.05$, fail to reject H_0

Lời giải chi tiết

$SE = \sqrt{\frac{36}{9} + \frac{30.25}{9}} = \sqrt{7.361} \approx 2.713$. $t = \frac{78 - 83}{2.713} \approx -1.84$. At $df = 8$: $t = 1.397$ has one-tail area 0.10; $t = 1.860$ has one-tail area 0.05. Since $1.397 < 1.84 < 1.860$, the two-tail P-value is between 0.10 and 0.20, so $P > 0.05$ and we fail to reject H_0. **(A)**.

Bài tập luyện tập

A horticulturist compares plant heights (cm) using two fertilizers on independent random samples. Fertilizer 1: $n_1 = 7$, $\bar{x}_1 = 32.4$, $s_1 = 4.2$. Fertilizer 2: $n_2 = 9$, $\bar{x}_2 = 28.1$, $s_2 = 3.6$ (both roughly symmetric, no outliers). Construct a 98% CI for $\mu_1 - \mu_2$.

Lời giải chi tiết

$df = \min(6, 8) = 6$, so a 98% CI needs one-tail area 0.01: $t^* = 3.143$. $SE = \sqrt{\frac{4.2^2}{7} + \frac{3.6^2}{9}} = \sqrt{2.52 + 1.44} = \sqrt{3.96} \approx 1.990$. $ME = 3.143(1.990) \approx 6.25$. CI: $(32.4 - 28.1) \pm 6.25 = 4.3 \pm 6.25 = (-1.95,\ 10.55)$ cm. The interval contains 0, so at the 98% level we cannot conclude a significant difference in mean plant height.

Bài tập luyện tập

A physical therapist measures the grip strength of the same 20 patients before and after a rehabilitation program. Which procedure is appropriate?

(A) Two-sample t procedure (independent)
(B) Matched-pairs t procedure
(C) Chi-square test
(D) One-proportion z procedure

Lời giải chi tiết

Each patient is measured twice (before and after), so the two sets of measurements are linked observation-by-observation – this calls for a matched-pairs t procedure on the within-patient differences. **(B)**.

Bài tập luyện tập

A coach wants to know if a new stretching routine improves 1-mile run times. Eight runners' times (minutes) are recorded before and after adopting the routine for 4 weeks; differences (before $-$ after) are: $0.4, 0.2, 0.5, -0.1, 0.3, 0.6, 0.2, 0.3$. Test at $\alpha = 0.01$ whether the routine reduces run time on average.

Lời giải chi tiết

State: $H_0 : \mu_d = 0$; $H_a : \mu_d > 0$, where μ_d is the true mean reduction (before $-$ after) in run time. $\alpha = 0.01$.
Plan: matched-pairs t-test, $df = 7$. Random, 10%, and Normal/Large Sample (assume no strong skew/outliers among the differences) are all satisfied.
Do: $\bar{d} = \dfrac{2.4}{8} = 0.3$. Sum of squared deviations from $\bar{d}$ is 0.32, so $s_d = \sqrt{0.32/7} \approx 0.2138$. $t = \dfrac{0.3}{0.2138/\sqrt{8}} = \dfrac{0.3}{0.0756} \approx 3.97$. At $df = 7$, $t = 3.499$ has one-tail area 0.005; since $3.97 > 3.499$, $P < 0.005$.
Conclude: Since $P < 0.005 < \alpha = 0.01$, we **reject** H_0: strong evidence the routine reduces mean run time.

Bài tập luyện tập

A tutor records SAT practice scores for $n = 6$ students before and after a 3-week prep course; differences (after $-$ before) are: $40, 60, 30, 50, 20, 60$. Construct a 90% confidence interval for the true mean score improvement μ_d.

Lời giải chi tiết

$\bar{d} = \dfrac{260}{6} \approx 43.33$. Sum of squared deviations from $\bar{d}$ is 1333.33, so $s_d^2 = 1333.33/5 = 266.67$, $s_d \approx 16.33$. $df = 5$, $t^* = 2.015$ (one-tail area 0.05 for 90% CI). $SE = \dfrac{16.33}{\sqrt{6}} \approx 6.667$. $ME = 2.015(6.667) \approx 13.43$. CI: $43.33 \pm 13.43 = (29.90,\ 56.76)$ points. We are 90% confident the true mean improvement is between 29.90 and 56.76 points.

Bài tập luyện tập

For a matched-pairs t procedure based on 14 pairs, the degrees of freedom equal:

(A) 13
(B) 14
(C) 26
(D) 12

Lời giải chi tiết

$df = n_{\text{pairs}} - 1 = 14 - 1 = 13$. **(A)**.

Bài tập luyện tập

Two graders (Grader X and Grader Y) each independently score the same 10 essays out of 100 points. Differences (X $-$ Y) are: $3, -2, 4, 1, 0, 5, -1, 2, 3, 1$. Decide whether a two-sample or matched-pairs procedure is appropriate, then construct a 95% CI for the mean difference in scoring.

Lời giải chi tiết

Decision: the same 10 essays are scored by both graders, so the scores are naturally paired by essay – a **matched-pairs** procedure is appropriate, not a two-sample procedure.
State: parameter is μ_d, the true mean difference (X $-$ Y) in scoring. Target: 95% CI.
Plan: matched-pairs t interval, $df = 9$. Random, 10%, Normal/Large Sample (assume no strong skew/outliers) are satisfied.
Do: $\bar{d} = \dfrac{16}{10} = 1.6$. Sum of squared deviations from $\bar{d}$ is 44.40, so $s_d^2 = 44.40/9 = 4.933$, $s_d \approx 2.221$. $t^* = 2.262$ ($df = 9$, one-tail area 0.025). $SE = \dfrac{2.221}{\sqrt{10}} \approx 0.702$. $ME = 2.262(0.702) \approx 1.59$.

$$CI : 1.6 \pm 1.59 = (0.01,\ 3.19) \text{ points.}$$

Conclude: We are 95% confident the true mean difference in scoring is between 0.01 and 3.19 points. Because the interval barely excludes 0, there is weak but real evidence that Grader X scores slightly higher on average than Grader Y.

Bài tập luyện tập

If the sample size for a t confidence interval is quadrupled (all else equal), the margin of error will approximately:

(A) double
(B) halve
(C) quarter
(D) stay the same

Lời giải chi tiết

$ME \propto \dfrac{1}{\sqrt{n}}$, so quadrupling n multiplies $\dfrac{1}{\sqrt{n}}$ by $\dfrac{1}{2}$, roughly halving the margin of error (a small additional decrease comes from t^* shrinking slightly as df grows). **(B)**.

Bài tập luyện tập

Pilot data suggest $s \approx 15$. Researchers want a 95% confidence interval with margin of error no more than 4. Using $z^* \approx 1.96$ as an approximation, find the minimum required sample size.

Lời giải chi tiết

$n \geq \left(\dfrac{1.96(15)}{4}\right)^2 = \left(\dfrac{29.4}{4}\right)^2 = (7.35)^2 = 54.02$. Since n must be a whole number and larger n only tightens precision, round up: $n = 55$.

Bài tập luyện tập

A researcher knows the population standard deviation σ exactly (an unusual situation) and wants a confidence interval for μ. Which distribution should be used?

(A) The t-distribution
(B) The standard Normal (z) distribution
(C) The chi-square distribution
(D) The F-distribution

Lời giải chi tiết

The t-distribution exists specifically to handle the extra uncertainty from estimating σ with s. If σ is actually known, that extra uncertainty does not exist, so the ordinary z procedure applies. **(B)**.

Bài tập luyện tập

A teacher wants to know if her single class's mean test score differs from the national average of 75 (a fixed, known value). She has scores from her one class ($n = 28$). Which procedure applies?

(A) One-sample t-test
(B) Two-sample t-test
(C) Matched-pairs t-test
(D) Chi-square test

Lời giải chi tiết

There is only one sample being compared to a fixed hypothesized value (75), not to a second sample – this is a one-sample t-test. **(A)**.

Bài tập luyện tập

An airline's stated checked-bag weight limit is 23 kg; the airline believes the true mean weight of bags checked does not exceed this. A random sample of $n = 25$ bags gives $\bar{x} = 23.8$ kg, $s = 2.5$ kg (roughly symmetric, no outliers). Test at $\alpha = 0.01$ whether the mean weight exceeds the limit.

Lời giải chi tiết

State: $H_0 : \mu = 23$; $H_a : \mu > 23$ kg. $\alpha = 0.01$.
Plan: one-sample t-test, $df = 24$. Random, 10%, Normal/Large Sample (roughly symmetric, no outliers) are satisfied.
Do: $t = \dfrac{23.8 - 23}{2.5/\sqrt{25}} = \dfrac{0.8}{0.5} = 1.60$. At $df = 24$: $t = 1.318$ has one-tail area 0.10; $t = 1.711$ has one-tail area 0.05. Since $1.318 < 1.60 < 1.711$, P is between 0.05 and 0.10.
Conclude: Since $P > 0.05 > \alpha = 0.01$, we **fail to reject** H_0: not enough evidence at the 1% level that mean checked-bag weight exceeds 23 kg.

Bài tập luyện tập

Statistical software reports $df = 18.6$ (a non-integer value) for a two-sample t-test, while a hand calculation using the conservative method gives $df = 11$. Explain why the two values differ, and which one is generally more precise.

Lời giải chi tiết

Software typically uses the **Welch–Satterthwaite** approximation, a formula that blends n_1, n_2, s_1, and s_2 to produce a more precise (and usually larger, non-integer) effective df, rather than simply taking $\min(n_1 - 1, n_2 - 1)$. The Welch df is generally more accurate, giving a narrower CI and a more precise P-value; the conservative $df = \min(n_1 - 1, n_2 - 1)$ used for hand calculations is a safe lower bound that never overstates the evidence, but is somewhat less precise.

Bài tập luyện tập

A study pairs each of 15 sets of identical twins, randomly assigning one twin in each pair to Diet A and the other to Diet B, then compares weight change after 10 weeks. Is this a two-sample or matched-pairs situation? Explain.

Lời giải chi tiết

This is a **matched-pairs** situation. Although two different diets are being compared (not the same subject measured twice), each pair of twins acts as a natural block: the twins share genetics and much of their environment, so pairing controls for those shared factors. The correct analysis reduces each twin pair to a single difference $d_i = (\text{Diet A weight change}) - (\text{Diet B weight change})$ and runs a one-sample t procedure on those 15 differences, with $df = 14$.

Bài tập luyện tập

A random sample of $n = 40$ delivery times (minutes) for a food-delivery app has $\bar{x} = 28.6$, $s = 6.4$. The distribution is strongly right-skewed with no extreme outliers. Construct a 95% confidence interval for the true mean delivery time μ, and justify that conditions are met.

Lời giải chi tiết

Conditions: Random: assume stated. 10%: satisfied. Normal/Large Sample: even though the distribution is strongly skewed, $n = 40 \geq 30$, so the Central Limit Theorem makes the Normal/Large Sample condition acceptable.

$df = 39$ (use $df = 40$ row, the closest tabulated value): $t^* \approx 2.021$. $SE = \dfrac{6.4}{\sqrt{40}} \approx 1.012$. $ME \approx 2.021(1.012) \approx 2.05$. CI: $28.6 \pm 2.05 = (26.55,\ 30.65)$ minutes. We are 95% confident the true mean delivery time is between 26.55 and 30.65 minutes.

Bài tập luyện tập

A random sample of $n = 13$ measurements of dissolved oxygen (mg/L) in a lake gives $\bar{x} = 7.9$, $s = 1.1$, with a dotplot showing a roughly symmetric shape and no outliers. An environmental standard requires a mean of at least 8.0 mg/L. Test at $\alpha = 0.10$ whether there is evidence the true mean is below the standard.

Lời giải chi tiết

State: $H_0 : \mu = 8.0$; $H_a : \mu < 8.0$ mg/L. $\alpha = 0.10$.
Plan: one-sample t-test, $df = 12$. Random, 10%, Normal/Large Sample (roughly symmetric, no outliers) are satisfied.

Do: $t = \dfrac{7.9 - 8.0}{1.1/\sqrt{13}} = \dfrac{-0.1}{0.3051} \approx -0.33$. This is far smaller in magnitude than even the $df = 12$, one-tail-area-0.10 value of 1.356, so $P > 0.10$.
Conclude: Since $P > 0.10 = \alpha$, we **fail to reject** H_0: not enough evidence that the true mean dissolved oxygen level is below the standard.

Inference for Categorical Data: Chi-Square

Mục tiêu Unit: Kiểm định khớp tốt (goodness of fit), kiểm định tính độc lập (independence) và tính đồng nhất (homogeneity) với thống kê χ^2; khung suy luận bốn bước State–Plan–Do–Conclude áp dụng cho χ^2; hình dạng và bậc tự do của phân phối χ^2; điều kiện Random, 10%, và Expected Counts ≥ 5 (vì sao dùng expected chứ không phải observed); phân tích đóng góp từng thành phần (component contribution); phân biệt ba loại kiểm định qua tình huống thực tế; và cách xử lý khi vi phạm điều kiện expected counts bằng cách gộp nhóm.

8.1 Review: the four-step inference framework

Every significance test introduced earlier in this course — for a single proportion, a difference of proportions, a single mean, or a difference of means — follows the same logical skeleton. The three chi-square tests in this unit follow it too; only the details of **State** and **Do** change. On the AP Statistics exam, a chi-square procedure is one of the required inference procedures students must be able to select correctly from a described scenario, and free-response questions built around it often also ask for a supporting graph or numerical summary before the formal test itself.

The four steps applied to chi-square tests

- **State:** identify the population(s) and the categorical variable(s) involved; write H_0 and H_a in words and, where useful, in symbols; give the significance level α. For every chi-square test, H_0 is a claim about a *distribution* of a categorical variable (never about a mean or a single proportion).
- **Plan:** name which of the three tests applies (goodness of fit, independence, or homogeneity — Sections 3–5 give the criteria) and verify the conditions common to all three: **Random**, **10%** (if sampling without replacement), and all **expected** counts ≥ 5.
- **Do:** compute the expected counts, the test statistic $\chi^2 = \sum \frac{(O-E)^2}{E}$, and the degrees of freedom; locate the P-value (or bracket it) using a chi-square table or technology.
- **Conclude:** compare P to α (equivalently, compare χ^2 to the critical value from the table) and state a decision about H_0 *in the context of the problem*, not just "reject" or "fail to reject."

VN **GIẢI THÍCH TIẾNG VIỆT**

Mọi kiểm định ý nghĩa thống kê (significance test) trong chương trình đều theo đúng bốn bước: **State** (nêu tổng thể, biến, giả thuyết H_0/H_a, mức α); **Plan** (chọn đúng loại kiểm định và kiểm tra điều kiện); **Do** (tính toán thống kê kiểm định và bậc tự do, tìm P-value); **Conclude** (so sánh với α và kết luận bằng ngôn ngữ của bài toán thực tế). Với các kiểm định χ^2, H_0 luôn là một phát biểu về **phân phối** của biến định tính — không phải về trung bình hay một tỉ lệ đơn lẻ.

8.1.1 The State step, side by side

The three chi-square tests differ mainly in how H_0 is phrased. Comparing the three wordings side by side helps prevent the single most common State-step error on this unit — writing a goodness-of-fit hypothesis when the sampling design actually calls for independence or homogeneity, or vice versa.

Test	Typical H_0 wording
Goodness of fit	"The distribution of [variable] in the population is [stated proportions]."
Independence	"[Variable 1] and [Variable 2] are independent in the population" (equivalently, "there is no association between [Variable 1] and [Variable 2]").
Homogeneity	"The distribution of [variable] is the same across [the groups/populations]."

Unlike a z-test or t-test for a proportion or mean, a chi-square test is **never two-sided or lower-tailed**. Because every term $(O-E)^2/E$ is non-negative, a large χ^2 statistic can only mean the data deviate *more* from H_0 than expected — small deviations from H_0 always give small (near-zero) values of χ^2. So even though H_a is worded as a non-directional statement ("not equally likely," "associated," "distributions differ"), the P-value is always the area to the *right* of the observed χ^2 under the chi-square curve.

Chi-square tests appear throughout the sciences and in industry wherever a categorical variable needs to be checked against a model: geneticists check offspring ratios against Mendelian predictions, election analysts check whether a poll's demographic breakdown matches the census, quality-control engineers check whether defect types occur in the proportions a process is supposed to produce, and social scientists check whether two categorical survey items are related. All of these are instances of one of the three tests developed in Sections 3–5.

Graphing calculators and statistical software compute χ^2, df, and an exact P-value directly (for example, a TI-84's χ^2`GOF-Test` for goodness of fit, or χ^2`2way-Test`, entered from a matrix of counts, for independence and homogeneity). For a GOF test, the observed counts go into one list and the hypothesized proportions (or the expected counts directly) into another; for independence/homogeneity, the counts are entered as a matrix matching the shape of the table, and the calculator returns χ^2, df, P, and the full matrix of expected counts in one step — useful for double-checking the Large Counts condition. When only a printed critical-value table is available, bracket the P-value between the two columns the computed χ^2 falls between, or simply compare χ^2 directly to the single critical value for the stated α — both approaches lead to the same reject/fail-to-reject decision.

8.2 The chi-square statistic and its distribution (χ^2)

8.2.1 Why we square the deviations

For a categorical variable sorted into k categories, let O denote an observed count and E the count expected if H_0 were exactly true. The chi-square statistic combines the mismatch in every category into a single number:

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

A signed sum $\sum(O-E)$ would be useless as a test statistic: because expected counts are built from the same totals as the observed counts (the same n, and, for independence/homogeneity, the same marginal totals), the positive and negative deviations *always* cancel exactly, giving $\sum(O-E)=0$ for every table, whether H_0 is a perfect description of the data or a terrible one. Squaring each

deviation removes the sign, so genuine mismatches accumulate instead of canceling, and it penalizes large deviations disproportionately — a deviation twice as large contributes four times as much to χ^2. Dividing by E then rescales each squared deviation *relative to the size of the expected count*: a raw miss of 5 candies is a much bigger relative surprise when $E = 10$ than when $E = 500$.

Ví dụ mẫu · Why raw deviations always cancel

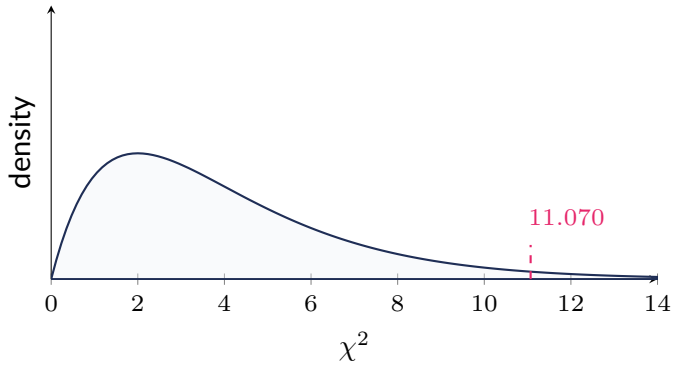
A variable has four categories with expected counts $E = 20, 20, 20, 20$ (total 80) and observed counts $O = 15, 25, 30, 10$ (total 80). The raw deviations are $-5, 5, 10, -10$, and indeed $\sum(O - E) = -5 + 5 + 10 - 10 = 0$ — no information about the mismatch survives. Squaring gives $25, 25, 100, 100$; dividing by $E = 20$ in each category gives components $1.25, 1.25, 5.0, 5.0$, so

$$\chi^2 = 1.25 + 1.25 + 5.0 + 5.0 = 12.5.$$

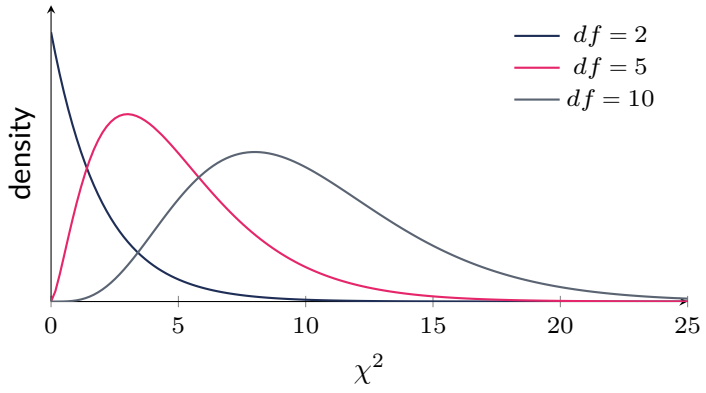
The squared-and-standardized version reveals exactly what the signed sum hid: categories 3 and 4 are the biggest sources of mismatch.

8.2.2 Shape of the chi-square distribution

The sampling distribution of χ^2 under H_0 is a family of curves indexed by degrees of freedom (df). Every member of the family shares three properties: it is defined only for $\chi^2 \geq 0$ (a sum of squared terms can never be negative), it is **right-skewed**, and its exact shape depends on df. As df increases, the distribution becomes taller in the middle, less skewed, and increasingly **symmetric and bell-shaped** — a consequence of the Central Limit Theorem, since a chi-square variable with $df = k$ is itself a sum of k independent squared standard-normal variables. The mean of a chi-square distribution equals its df; the variance equals $2\,df$. It is worth remembering that the chi-square curve is itself only an *approximation* to the true, discrete sampling distribution of $\sum(O-E)^2/E$ — an approximation that becomes accurate as sample sizes grow, which is exactly why the Large Counts condition (discussed next) exists: it is the practical rule of thumb that keeps this approximation trustworthy.



Chi-square distribution ($df = 4$, shown for shape): right-skewed, always non-negative; the critical value 11.070 ($df = 5$, $\alpha = 0.05$) is marked for reference.



Three members of the chi-square family. At $df = 2$ the curve is monotonically decreasing from its peak at 0; at $df = 5$ it peaks away from 0 but is still clearly right-skewed; at $df = 10$ it is markedly more symmetric and bell-shaped, and its peak has shifted rightward toward its mean of 10.

VN **GIẢI THÍCH TIẾNG VIỆT**

Phân phối χ^2 luôn không âm ($\chi^2 \geq 0$) và **lệch phải** (right-skewed), vì nó là tổng của các số hạng đã bình phương. Hình dạng cụ thể phụ thuộc vào **bậc tự do** (df): df càng lớn, đường cong càng **đối xứng và có dạng hình chuông**, đỉnh càng dịch sang phải (vì trung bình của phân phối χ^2 chính bằng df). Đây là lý do các bảng giá trị tới hạn (critical value) của χ^2 luôn được lập riêng cho từng df.

The chi-square goodness-of-fit test was introduced by the statistician Karl Pearson in 1900, making it one of the earliest formal hypothesis-testing procedures in statistics — decades before the four-step State/Plan/Do/Conclude framework used throughout this course was standardized for teaching. The tests for independence and homogeneity are later extensions of Pearson's original idea to two-way tables. Pearson's key insight was that the sum $\sum(O-E)^2/E$, computed from any large random sample, behaves approximately like a known mathematical curve (the chi-square distribution) regardless of the specific categories or context involved — which is precisely what makes a single family of critical-value tables usable across genetics, manufacturing, polling, and every other field where categorical data arise.

8.2.3 Degrees of freedom: where the formulas come from

The degrees of freedom of a chi-square test count how many of the expected counts are genuinely free to vary once the necessary totals are held fixed. In a goodness-of-fit test with k categories, the total n is fixed, so once $k-1$ of the expected counts are chosen the last one is completely determined (all counts must sum to n) — exactly $k-1$ quantities are free, hence $df = k-1$. In a two-way table with r rows and c columns, fixing every row total and every column total imposes $r+c-1$ independent constraints on the table (one constraint is redundant, because the row totals and the column totals both have to add up to the same grand total), which leaves $(r-1)(c-1)$ cells free to vary before the remaining cells are forced by the fixed totals — hence $df = (r-1)(c-1)$ for both the independence and the homogeneity tests.

8.2.4 Reading a chi-square critical-value table

A chi-square table lists critical values indexed by df (rows) and upper-tail area α (columns). To use it: locate the row for the test's df, then compare the computed χ^2 to the entry in the column for the chosen α. If χ^2 exceeds that entry, reject H_0 at that α. If χ^2 falls between two columns, the P-value can be bracketed between the corresponding α values — for example, a χ^2 that falls between the $\alpha = 0.05$ and $\alpha = 0.025$ columns means $0.025 < P < 0.05$. As a further illustration, a test with $df = 8$ and $\chi^2 = 17.0$ falls between the $\alpha = 0.05$ entry (15.507) and the $\alpha = 0.025$ entry (17.535) in the row for $df = 8$, so $0.025 < P < 0.05$ without needing any further computation.

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
1	2.706	3.841	5.024	6.635
2	4.605	5.991	7.378	9.210
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.070	12.833	15.086
6	10.645	12.592	14.449	16.812
8	13.362	15.507	17.535	20.090
10	15.987	18.307	20.483	23.209
12	18.549	21.026	23.337	26.217

Selected chi-square critical values used throughout this unit (a full table also lists every intermediate df).

8.2.5 Conditions common to all three chi-square tests: Random, 10%, and Expected Counts ≥ 5

Every chi-square procedure in this unit — goodness of fit, independence, homogeneity — requires the same three conditions before the chi-square curve can be trusted as an approximation to the true sampling distribution of the test statistic.

> **Random:** the data come from a random sample or a randomized experiment, so the results are not systematically biased and can be generalized to the population(s) of interest.
>
> **10% condition:** if sampling *without replacement* from a finite population, the sample size should be no more than 10% of the population, so that individual observations are close to independent.
>
> **Large Counts (Expected Counts):** *every* expected count $E_i \geq 5$.

The Large Counts condition is stated in terms of **expected**, not observed, counts, and the distinction matters. The chi-square distribution is a continuous curve used to approximate the true (discrete) sampling distribution of χ^2 under repeated random sampling from H_0. That approximation is only accurate when every category's expected count is reasonably large; when an expected count is small, the possible values a count can actually take are sparse and lumpy, and the smooth chi-square curve badly misrepresents the true P-value (typically making it too easy to reject a true H_0). This is purely a statement about *what* H_0 *predicts*, which is exactly what expected counts represent — it says nothing about what a particular sample happens to show. A cell with a small *observed* count next to a large expected count is not a violation of anything; it is simply evidence against H_0, which is exactly what the test is designed to detect. A cell with a small *expected* count, on the other hand, is a structural problem with the approximation itself, regardless of what is observed there.

For instance, suppose a goodness-of-fit test with $n = 40$ has just two categories with hypothesized proportions $p = 0.95$ and $p = 0.05$, giving expected counts 38 and 2. Even if the observed count in the small category turns out to be 10 — a striking, interesting deviation from an expected count of only 2 — the Large Counts condition is still violated, because it is the *expected* count of 2, not the observed count of 10, that determines whether the chi-square curve can be trusted as an approximation.

> (!) Lỗi thường gặp: **Không dùng % để tính χ^2.** Thống kê χ^2 luôn tính trên **số đếm** (counts), **không** dùng phần trăm hay tỉ lệ. Ngoài ra, điều kiện ≥ 5 áp dụng cho **expected counts** chứ **không phải** observed counts — một ô có observed nhỏ nhưng expected lớn vẫn hoàn toàn hợp lệ (đó chính là bằng chứng chống lại H_0); ngược lại, nếu *expected* < 5 ở bất kỳ ô nào, điều kiện bị vi phạm và cần gộp nhóm hoặc thu thêm dữ liệu trước khi kết luận (xem Section 7).

> (!) Lỗi thường gặp: **"Independence" mang hai nghĩa khác nhau — đừng nhầm lẫn.** Điều kiện **Random** ngầm giả định các **quan sát độc lập với nhau** (observations are independent) — mỗi cá thể không ảnh hưởng đến xác suất xảy ra của cá thể khác — và điều kiện này áp dụng cho *cả ba* loại kiểm định χ^2. Khái niệm này hoàn toàn khác với tên gọi **"kiểm định tính độc lập" (test for independence)** ở Section 4, vốn nói về việc *hai biến định tính* có liên hệ với nhau hay không. Một kiểm định goodness-of-fit hay homogeneity vẫn cần "independent observations" dù nó không phải là "test for independence."

8.3 Chi-square goodness-of-fit test (χ^2 GOF)

A chi-square goodness-of-fit (GOF) test uses *one* random sample from a single population and asks whether a single categorical variable follows a specific hypothesized distribution. Typical uses include checking a genetic cross against a Mendelian ratio, checking whether a manufacturing process produces defect types in their historical proportions, and checking whether a random-number generator or a die produces outcomes that are actually equally likely.

H_0: the categorical variable follows the stated population proportions $p_1, p_2, \dots, p_k$. H_a: the variable does *not* follow that distribution (at least one p_i is wrong).

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \qquad E_i = np_i, \qquad df = k - 1.$$

VN **GIẢI THÍCH TIẾNG VIỆT**

Kiểm định khớp tốt (goodness of fit) chỉ dùng cho **MỘT** biến định tính trên **MỘT** mẫu, so sánh phân phối quan sát được với một phân phối **giả thuyết đã cho trước** (có thể là đều nhau — equally likely — hoặc không đều nhau, miễn là các tỉ lệ p_i được nêu rõ và cộng lại bằng 100%). Bậc tự do luôn bằng số nhóm trừ 1.

8.3.1 Equal proportions: a uniform hypothesized distribution

Ví dụ mẫu · Candy colors — equally likely?

A bag of 60 candies should have colors in equal proportion (six colors). Observed counts: $8, 12, 9, 11, 10, 10$. Test H_0 : colors are equally likely vs. H_a : they are not, at $\alpha = 0.05$.
State: Let $p_1, \dots, p_6$ be the population proportions of the six colors. $H_0 : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$. H_a : at least one $p_i \neq \frac{1}{6}$. $\alpha = 0.05$.
Plan: Chi-square goodness-of-fit test, $df = 6 - 1 = 5$. *Random:* assume the candies in the bag are randomly mixed. *10%:* 60 candies are far fewer than 10% of all candies the manufacturer produces. *Expected counts:* each expected count is $60/6 = 10 \geq 5$ — condition met.
Do: Expected count for each color $= 60/6 = 10$. The squared deviations $(O - E)^2$ for the six colors are $4, 4, 1, 1, 0, 0$, and every expected count is 10, so

$$\chi^2 = \frac{4 + 4 + 1 + 1 + 0 + 0}{10} = \frac{10}{10} = 1.0.$$

Conclude: With $df = 5$, the critical value at $\alpha = 0.05$ is 11.070. Since $1.0 < 11.070$ (in fact χ^2 is even below the $\alpha = 0.10$ critical value of 9.236, so $P > 0.10$), **fail to reject** H_0: the data are consistent with an equal-proportion distribution of colors.

8.3.2 Unequal proportions: a non-uniform hypothesized distribution

Not every GOF hypothesis states that categories are equally likely; H_0 may specify any set of proportions that sum to 100%, and each category then gets its own expected count $E_i = np_i$.

Ví dụ mẫu · Testing a manufacturer's stated color mix

A candy manufacturer states that its assorted-color candies are produced in the long-run proportions Blue 24%, Orange 20%, Green 16%, Yellow 14%, Red 13%, Brown 13%. A quality-control inspector randomly samples $n = 200$ candies from a shipment and records the color counts below. Test whether the shipment's color composition is consistent with the manufacturer's stated distribution, at $\alpha = 0.05$.

Color	Stated %	Observed O	Expected $E = np_i$
Blue	24%	66	48
Orange	20%	30	40
Green	16%	22	32
Yellow	14%	32	28
Red	13%	16	26
Brown	13%	34	26
Total	100%	200	200

State: H_0: the shipment's color proportions match the stated distribution $(24\%, 20\%, 16\%, 14\%, 13\%, 13\%)$. H_a: at least one color proportion differs from the stated value. $\alpha = 0.05$.
Plan: Chi-square goodness-of-fit test, $df = 6 - 1 = 5$. *Random:* the 200 candies were sampled randomly from the shipment. *10%:* 200 candies are less than 10% of the full shipment. *Expected counts:* from the table, the smallest expected count is $26 \geq 5$ — condition met.
Do:

$$\chi^2 = \frac{18^2}{48} + \frac{(-10)^2}{40} + \frac{(-10)^2}{32} + \frac{4^2}{28} + \frac{(-10)^2}{26} + \frac{8^2}{26} \approx 6.75 + 2.5 + 3.125 + 0.571 + 3.846 + 2.462 \approx 19.25.$$

Conclude: With $df = 5$, the critical value at $\alpha = 0.05$ is 11.070, and $19.25 > 11.070$; in fact 19.25 also exceeds the $\alpha = 0.01$ critical value of 15.086, so $P < 0.01$. **Reject** H_0: there is convincing evidence that the shipment's color composition differs from the manufacturer's stated distribution. In practice, a result this strong would likely prompt the inspector to flag the shipment for further investigation, since a systematic shift this large across several colors more plausibly reflects a problem upstream (for example, a miscalibrated color-mixing step) than ordinary sampling variation.

VN **GIẢI THÍCH TIẾNG VIỆT**
Khi giả thuyết không phải "đều nhau", mỗi nhóm có expected count riêng $E_i = n \times p_i$ (với p_i là tỉ lệ giả thuyết của nhóm đó), thay vì chia đều n/k. Cách tính χ^2, $df = k - 1$, và cách so sánh với giá trị tới hạn hoàn toàn giống trường hợp "đều nhau" — chỉ khác ở bước tính E_i.

8.3.3 Reading component contributions to χ^2

Each term $(O_i - E_i)^2/E_i$ is that category's **component** of the total statistic — it measures how much that single category contributed to the evidence against H_0. Reading the components lets you go beyond "reject/fail to reject" and explain *which* categories drove the result.

In the color example above, the six components were Blue 6.75, Orange 2.5, Green 3.125, Yellow 0.571, Red 3.846, Brown 2.462. Blue and Red together contribute $6.75 + 3.846 = 10.60$, roughly 55% of the total $\chi^2 = 19.25$ — the shipment contains noticeably more blue candies and noticeably fewer red candies than the manufacturer's stated distribution predicts. Yellow's component (0.571, under 3% of the total) shows that color came out very close to its expected count and contributed almost nothing to the rejection of H_0.

(!) Lỗi thường gặp: Phân tích component chỉ mang tính **mô tả** (descriptive), không phải một kiểm định riêng biệt. Không được chọn ra một component lớn nhất rồi tuyên bố nhóm đó "có ý nghĩa thống kê" một cách độc lập — kết luận reject/fail to reject chỉ áp dụng cho χ^2 **tổng**, không áp dụng cho từng ô riêng lẻ.

8.3.4 Reading exact P-values from technology

Statistical software, or a calculator's built-in goodness-of-fit test, reports an exact P-value instead of only a table-based bracket. The next example shows how that exact value should relate to the critical-value table.

Ví dụ mẫu · Customer complaint categories

A company's complaint log has historically shown Billing 30%, Shipping 25%, Product Quality 20%, Customer Service 25% of all complaints. This month, a random sample of $n = 180$ complaints gives Billing 62, Shipping 40, Quality 30, Service 48. Test whether this month's pattern differs from the historical distribution, at $\alpha = 0.05$.

State: H_0: this month's complaint distribution matches the historical distribution ($30\%, 25\%, 20\%, 25\%$). H_a: at least one category's proportion has changed. $\alpha = 0.05$.

Plan: Chi-square goodness-of-fit test, $df = 4 - 1 = 3$. *Random:* the 180 complaints are treated as a random sample of this month's complaints. *10%:* 180 complaints are a small fraction of all complaints the company could ever receive. *Expected counts:* $0.30(180) = 54$, $0.25(180) = 45$, $0.20(180) = 36$, $0.25(180) = 45$ — all ≥ 5, condition met.

Do: $\chi^2 = \dfrac{8^2}{54} + \dfrac{(-5)^2}{45} + \dfrac{(-6)^2}{36} + \dfrac{3^2}{45} = 1.185 + 0.556 + 1.000 + 0.200 \approx 2.94$.

Conclude: With $df = 3$, the critical value at $\alpha = 0.05$ is 7.815, and even the more lenient $\alpha = 0.10$ critical value is only 6.251. Since 2.94 is well below both, $P > 0.10$; a calculator's exact P-value output would confirm this directly rather than requiring a bracket. **Fail to reject** H_0: this month's complaint pattern is consistent with the company's historical distribution.

8.4 Chi-square test for independence

A chi-square test for independence uses *one* random sample from a single population and classifies each individual on *two* categorical variables, asking whether the two variables are associated. Market researchers use it to check whether product preference is related to a customer's region; public-health researchers use it to check whether a risk factor is related to an outcome; pollsters use it to check whether opinion on an issue is related to demographic category — in every case, a single sample is cross-classified on two variables at once.

H_0: the two categorical variables are independent in the population (no association). H_a: the two variables are associated (not independent).

$$E_{ij} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{n}, \qquad df = (r-1)(c-1).$$

VN **GIẢI THÍCH TIẾNG VIỆT**

Kiểm định tính độc lập (independence) dùng **MỘT** tổng thể duy nhất, mỗi cá thể được phân loại theo **HAI** biến định tính cùng lúc, và câu hỏi là hai biến đó có **liên hệ** (associated) với nhau hay không. Expected count của mỗi ô $= \dfrac{\text{tổng hàng} \times \text{tổng cột}}{\text{tổng chung}}$; $df = (r-1)(c-1)$ với r, c là số hàng, số cột.

8.4.1 Example: gender and subject preference

Ví dụ mẫu · Gender and subject preference

Is there an association between gender and preferred subject? A random sample of $n = 120$ students gives the table below. Test at $\alpha = 0.05$.

	Math	Science	Arts	Total
Male	30	20	10	60
Female	20	25	15	60
Total	50	45	25	120

State: H_0: gender and subject preference are independent. H_a: gender and subject preference are associated. $\alpha = 0.05$.
Plan: Chi-square test for independence, $df = (2-1)(3-1) = 2$. *Random:* the 120 students are a random sample. *10%:* 120 students are less than 10% of all students at the relevant population level. *Expected counts:* computed below, all ≥ 5 — condition met.
Do: Expected counts (row total$\times$col total$/120$): Male–Math 25, Male–Science 22.5, Male–Arts 12.5, Female–Math 25, Female–Science 22.5, Female–Arts 12.5.

$$\chi^2 = \frac{5^2}{25} + \frac{2.5^2}{22.5} + \frac{2.5^2}{12.5} + \frac{5^2}{25} + \frac{2.5^2}{22.5} + \frac{2.5^2}{12.5} = 1 + 0.278 + 0.5 + 1 + 0.278 + 0.5 \approx 3.56.$$

Conclude: With $df = 2$, the critical value at $\alpha = 0.05$ is 5.991. Since $3.56 < 5.991$, **fail to reject** H_0: there is no significant evidence of an association between gender and subject preference.

8.4.2 Visualizing association before testing: conditional distributions

Before running a formal test, it is good practice to describe what a two-way table shows using **conditional distributions** — the percentage breakdown of the column variable computed separately within each row. For the gender/subject table above:

	Math	Science	Arts
Male (of $n = 60$ males)	50.0%	33.3%	16.7%
Female (of $n = 60$ females)	33.3%	41.7%	25.0%

The two conditional distributions look different — males lean more toward Math, females more toward Science and Arts — which is exactly the kind of pattern that motivates running a formal test in the first place. But in this particular sample, the gap between the two rows ($\chi^2 = 3.56$ against a critical value of 5.991) turned out to be small enough to be plausibly explained by chance sampling variation alone. A chi-square test is precisely what distinguishes a real difference between conditional distributions from the ordinary sample-to-sample variation any random sample would show even when H_0 is exactly true.

8.4.3 Example: age group and preferred news source

The gender/subject example above did not reject H_0. The next example shows the same mechanics leading to the opposite conclusion.

Ví dụ mẫu · Age group and news source — a significant association

A random sample of $n = 300$ adults is classified by age group and by the news source they use most often. Test whether age group and news source are associated, at $\alpha = 0.05$.

Age group	Social media	TV	Newspaper	Total
18–34	70	20	10	100
35–54	35	40	25	100
55+	15	40	45	100
Total	120	100	80	300

State: H_0: age group and preferred news source are independent. H_a: age group and preferred news source are associated. $\alpha = 0.05$.
Plan: Chi-square test for independence, $df = (3-1)(3-1) = 4$. *Random:* the 300 adults form a random sample. *10%:* 300 adults are less than 10% of the adult population. *Expected counts:* shown in the table below (row total$\times$col total$/300$); the smallest is $26.67 \geq 5$ — condition met.

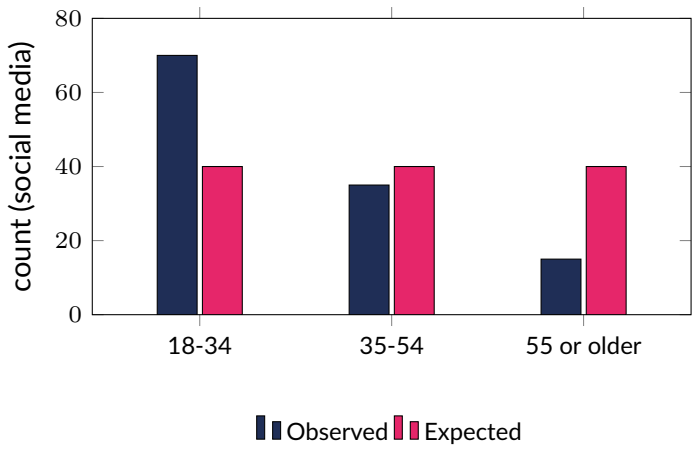
Age group	Social media	TV	Newspaper
18–34	40.00	33.33	26.67
35–54	40.00	33.33	26.67
55+	40.00	33.33	26.67

Do:

$$\chi^2 = \frac{30^2}{40} + \frac{(-13.33)^2}{33.33} + \frac{(-16.67)^2}{26.67} + \frac{(-5)^2}{40} + \frac{6.67^2}{33.33} + \frac{(-1.67)^2}{26.67} + \frac{(-25)^2}{40} + \frac{6.67^2}{33.33} + \frac{18.33^2}{26.67}$$

$$\approx 22.50 + 5.33 + 10.42 + 0.63 + 1.33 + 0.10 + 15.63 + 1.33 + 12.60 \approx 69.87.$$

Conclude: With $df = 4$, the critical value at $\alpha = 0.05$ is 9.488, and $\chi^2 = 69.87$ is far larger — it also exceeds the $\alpha = 0.01$ critical value of 13.277, so $P < 0.01$. **Reject** H_0: there is convincing evidence that age group and preferred news source are associated. Younger adults in the sample favor social media far more than expected, while adults 55 and older favor newspapers far more than expected. A finding this strong would typically motivate follow-up research into why the age groups differ so much in their media habits — though, as Section 4.6 discusses, the test itself only establishes that an association exists, not why it exists.



Observed vs. expected counts choosing "social media" as their primary news source, by age group, from the age$\times$news-source example. The largest gaps between observed and expected bars occur for the youngest and oldest groups — consistent with the large χ^2 statistic and the decision to reject H_0.

8.4.4 Component contributions in a two-way table

Just as in a goodness-of-fit test, each cell of a two-way table contributes its own term $(O-E)^2/E$ to the total χ^2, and reading these components shows *which* cells drove a significant result.

In the age-group/news-source example, the nine components were (row by row, in the order they appear in the table): $22.50, 5.33, 10.42$ for ages 18–34; $0.63, 1.33, 0.10$ for ages 35–54; $15.63, 1.33, 12.60$ for ages 55+. The single largest component, 22.50, comes from the youngest group's social-media cell — that one cell alone accounts for roughly 32% of the entire test statistic ($22.50/69.87 \approx 0.32$). The oldest group's two large components (social media, 15.63; newspaper, 12.60) show that group deviating from expected counts in *both* directions at once — far below expected on social media, far above expected on newspaper. The middle-aged group, by contrast, contributed almost nothing ($0.63 + 1.33 + 0.10 = 2.06$, under 3% of the total): its news-source distribution came out close to what independence would predict.

8.4.5 The special case of a 2×2 table: connection to the two-proportion z-test

When a two-way table has exactly two rows and two columns, $df = (2-1)(2-1) = 1$, and the chi-square test becomes mathematically equivalent to the two-proportion z-test from an earlier unit: testing $H_0 : p_1 = p_2$ against a two-sided $H_a : p_1 \neq p_2$ satisfies $\chi^2 = z^2$, where z is the test statistic from the two-proportion z-test on the same data. In the car-ownership practice problem (Independence practice, below), $\chi^2 = 16.0$, so the equivalent two-proportion z-test on that same 2×2 table would give $z = \pm\sqrt{16.0} = \pm 4.0$ — the same conclusion, reached from an algebraically identical direction. This equivalence holds only for 2×2 tables ($df = 1$); it has no analogue once a table has more than two rows or columns, since there is no single "z" statistic once more than two proportions are being compared at once.

VN **GIẢI THÍCH TIẾNG VIỆT**

Với bảng 2×2 ($df = 1$), kiểm định χ^2 và kiểm định hai tỉ lệ (two-proportion z-test) **tương đương về mặt toán học**: $\chi^2 = z^2$. Đây cũng là lý do giá trị tới hạn χ^2 ở $df = 1$, $\alpha = 0.05$ (là 3.841) chính là bình phương của giá trị tới hạn $z = 1.96$ quen thuộc ($1.96^2 \approx 3.8416$). Quan hệ này **chỉ đúng** khi bảng có đúng 2 hàng và 2 cột — không áp dụng cho bảng lớn hơn.

8.4.6 What a significant chi-square result does — and does not — tell you

Three cautions apply to every chi-square conclusion, no matter which of the three tests produced it.

Association is not causation. Rejecting H_0 in a test for independence shows that two variables measured on the same observational sample are related, but it does not, by itself, show that one *causes* the other — a lurking variable could explain both. A chi-square result supports a causal claim only when the explanatory categorical variable was assigned to subjects by **random assignment** in a designed experiment, which is exactly the setting behind many homogeneity tests (subjects randomly assigned to, say, different treatments or different diets).

χ^2 measures the amount of evidence against H_0, not the size or direction of any relationship. A large χ^2 says the data are unlikely under H_0; on its own it says nothing about which categories are linked, how strongly, or in which direction. That information has to come from examining the actual counts, the component-contribution discussions earlier in Sections 3 and 4, or a graph of conditional distributions.

Statistical significance is not the same as practical importance. With a large enough sample, even a tiny, practically unimportant deviation from H_0 can produce a large χ^2 and a rejected null. For example, testing $H_0 : p = 0.50$ vs. $H_a : p \neq 0.50$ for a two-category variable with $n = 10{,}000$ and observed counts 5100 and 4900 (a shift of only one percentage point) gives expected counts of 5000 and 5000, so

$$\chi^2 = \frac{100^2}{5000} + \frac{(-100)^2}{5000} = 2.0 + 2.0 = 4.0,$$

which exceeds the $df = 1$, $\alpha = 0.05$ critical value of 3.841 — technically significant, even though a shift from 50% to 51% may be of little practical consequence in most real contexts. Always report and interpret the actual counts and proportions alongside the significance decision, not the decision alone.

VN **GIẢI THÍCH TIẾNG VIỆT**

Ba lưu ý quan trọng khi đọc kết quả χ^2: (1) **Có liên hệ không đồng nghĩa với có quan hệ nhân quả** (association is not causation) — chỉ khi biến giải thích được **gán ngẫu nhiên** (random assignment) trong một thí nghiệm mới có thể suy luận nhân quả từ kết quả có ý nghĩa thống kê. (2) χ^2 lớn chỉ cho biết **có bằng chứng** chống lại H_0, không cho biết mối liên hệ mạnh hay yếu, theo chiều nào — muốn biết điều đó phải xem lại bảng số liệu, các component, hoặc phân phối có điều kiện. (3) Với mẫu đủ lớn, ngay cả một khác biệt rất nhỏ và không có ý nghĩa thực tiễn cũng có thể cho ra kết quả "có ý nghĩa thống kê" (statistically significant) — luôn báo cáo số liệu thực tế đi kèm với kết luận, không chỉ riêng kết luận reject/fail to reject.

8.5 Chi-square test for homogeneity

Independence and homogeneity share identical mechanics — the same expected-count formula and the same $df = (r-1)(c-1)$ — but they arise from a different sampling design and answer a different question. A homogeneity test begins with *two or more separate* populations or groups, samples independently within each one, and classifies every individual on the *same one* categorical variable, asking whether the groups share the same distribution of that variable. Comparative clinical trials, comparisons of outcome rates across schools, hospitals, or factories, and before/after comparisons of an intervention on separate cohorts are all classic homogeneity settings.

H_0: the distribution of the categorical variable is the same across all populations/groups. H_a: at least one population/group has a different distribution. Expected counts and df are computed exactly as for independence, using the row and column totals of the combined table.

8.5.1 Example: comparing three weight-loss diets

Ví dụ mẫu · Comparing three weight-loss diets

Independent random samples of $n = 80$ patients are assigned to each of three diets (A, B, C), and each patient's outcome after eight weeks is recorded. Test whether the outcome distribution is the same across the three diets, at $\alpha = 0.05$.

Diet	Lost weight	No change	Gained weight	Total
A	45	20	15	80
B	30	25	25	80
C	25	20	35	80
Total	100	65	75	240

State: H_0: the distribution of weight-change outcome is the same for all three diets. H_a: the distribution of weight-change outcome differs for at least one diet. $\alpha = 0.05$.
Plan: Chi-square test for homogeneity, $df = (3-1)(3-1) = 4$. *Random:* patients were independently randomly assigned/sampled within each diet group. *10%:* each group of 80 is a small fraction of all possible patients who could follow that diet. *Expected counts:* shown below; the smallest is $21.67 \geq 5$ — condition met.

Diet	Lost weight	No change	Gained weight
A	33.33	21.67	25.00
B	33.33	21.67	25.00
C	33.33	21.67	25.00

Do: Row-by-row components $(O-E)^2/E$:

Diet A: $4.083 + 0.128 + 4.000 = 8.211$ Diet B: $0.333 + 0.513 + 0.000 = 0.846$

Diet C: $2.083 + 0.128 + 4.000 = 6.211$

$$\chi^2 = 8.211 + 0.846 + 6.211 \approx 15.27.$$

Conclude: With $df = 4$, the critical value at $\alpha = 0.05$ is 9.488, and the critical value at $\alpha = 0.01$ is 13.277. Since $15.27 > 13.277$, $P < 0.01$, so **reject** H_0: there is convincing evidence that the distribution of weight-change outcome is not the same across the three diets. Diet A produced far more "lost weight" results than expected under a common distribution ($O = 45$ vs. $E = 33.33$, component 4.08), while Diet C produced far more "gained weight" results than expected ($O = 35$ vs. $E = 25$, component 4.00); together these two cells account for over half of the total test statistic. Because patients were randomly assigned to diets in this design, this result — unlike a purely observational independence test — does support the causal claim that the choice of diet affects weight-change outcome, not merely that diet and outcome happen to be associated.

VN **GIẢI THÍCH TIẾNG VIỆT**

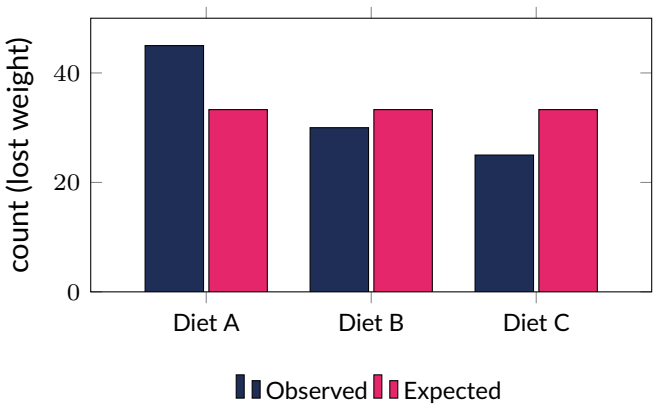
Kiểm định homogeneity dùng công thức và cách tính χ^2, df giống hệt independence — điểm khác biệt duy nhất nằm ở **thiết kế lấy mẫu**: homogeneity bắt đầu từ **nhiều tổng thể/nhóm riêng biệt** đã biết trước (ví dụ: ba nhóm được gán ngẫu nhiên vào ba chế độ ăn khác nhau), trong khi independence bắt đầu từ **một** mẫu ngẫu nhiên duy nhất rồi mới phân loại theo hai biến. Vì cơ chế tính toán giống nhau, một số sách gọi chung cả hai là "kiểm định hai chiều" (two-way table test).

8.5.2 Visualizing distributions before testing

As with independence, it helps to compare the conditional (row) distributions before formally testing them. For the three-diet table:

Diet	Lost weight	No change	Gained weight
A (of $n = 80$)	56.25%	25.00%	18.75%
B (of $n = 80$)	37.50%	31.25%	31.25%
C (of $n = 80$)	31.25%	25.00%	43.75%

Diet A's row is noticeably shifted toward "lost weight" compared to the other two, and Diet C's row is noticeably shifted toward "gained weight." These visual differences in the conditional distributions are consistent with — and foreshadow — the significant result the formal test produced: a common underlying distribution across all three diets would be expected to make the three rows of percentages look much more alike than they do here.



Observed vs. expected counts of patients who lost weight, by diet, from the three-diet homogeneity example. Diet A shows more "lost weight" patients than expected under a common distribution, and Diet C shows fewer.

8.6 Choosing the right chi-square test

The three tests use identical arithmetic once expected counts are found, so the decision of *which* test applies rests entirely on the sampling design described in the problem.

VN **GIẢI THÍCH TIẾNG VIỆT**

Phân biệt ba loại kiểm định χ^2: **Goodness of fit** — MỘT tổng thể, MỘT biến định tính, so với một phân phối **giả thuyết** cho trước; **Independence** — MỘT tổng thể, HAI biến định tính, hỏi có liên hệ không; **Homogeneity** — NHIỀU tổng thể/nhóm (đã lấy mẫu riêng biệt), MỘT biến định tính, hỏi phân phối có giống nhau giữa các nhóm không. Công thức tính toán (χ^2, expected, cách so sánh df) giống hệt nhau ở cả ba trường hợp; điểm khác biệt chỉ nằm ở **cách đặt câu hỏi** và **cách lấy mẫu** (bao nhiêu tổng thể, bao nhiêu biến).

Goodness of fit	Independence	Homogeneity
1 population, 1 categorical variable.	1 population, 2 categorical variables.	≥ 2 populations/groups, 1 categorical variable.
Compares observed counts to a *hypothesized* distribution.	Asks whether the two variables are *associated*.	Asks whether the groups share the *same distribution*.
$df = k - 1$	$df = (r-1)(c-1)$	$df = (r-1)(c-1)$

The three chi-square tests, distinguished by how many populations are sampled and how many categorical variables are recorded.

Scenario 1. A single random sample of 500 U.S. adults is classified by both political party (Democrat/Republican/Independent) *and* opinion on a policy proposal (Support/Oppose/Undecided). → **Independence:** one sample, two categorical variables recorded on the same individuals, asking whether party and opinion are associated.

Scenario 2. A biologist wants to know whether the wing-color distribution of a butterfly population matches the $9{:}3{:}3{:}1$ ratio predicted by a genetic model; a sample of 250 butterflies is classified by wing color. → **Goodness of fit:** one sample, one variable, compared against a distribution stated in advance by the genetic model.

Scenario 3. Independent random samples of 150 students are drawn separately from three different high schools, and each student is classified by preferred learning format (In-person/Online/Hybrid). → **Homogeneity:** three separate populations (the three schools) sampled independently, each classified on the same one variable, asking whether the format preference distribution is the same across schools.

Scenario 4. A national retailer wants to know whether the payment-method mix at its own stores (Card/Cash/Mobile) matches the industry-wide proportions reported in a trade publication

(55%, 25%, 20%). The retailer samples 300 transactions from its own stores only. → **Goodness of fit:** even though the comparison is being made *against* an external, publicly reported benchmark, there is still only one sample (the retailer's own transactions) and one categorical variable (payment method), tested against a distribution stated in advance — not a comparison of several separately sampled groups, so this is not homogeneity.

Test	Sampling design	H_0 compares	df
Goodness of fit	One random sample, one categorical variable	Observed distribution vs. a stated distribution	$k-1$
Independence	One random sample, two categorical variables	The two variables vs. no association between them	$(r-1)(c-1)$
Homogeneity	Independent samples from ≥ 2 groups, one categorical variable	The distributions across groups vs. one common distribution	$(r-1)(c-1)$

Side-by-side summary of the three chi-square procedures covered in this unit.

8.7 When conditions fail: combining categories

If any expected count in a chi-square table is below 5, the Large Counts condition is violated and the chi-square approximation cannot be trusted as stated. Two standard remedies exist: **collect more data** (a larger n raises every expected count proportionally) or, when that is not possible, **combine categories** that are conceptually similar so that every resulting expected count reaches 5 or more. Combining categories reduces the number of categories k (or the number of rows/columns), which correspondingly reduces df.

Combining categories is a practical fallback, not a first choice. Whenever collecting a larger sample is realistic, doing so is preferable: a larger n raises every expected count in proportion without discarding any of the original category structure or the detail it carries. Combining is reserved for situations where some categories are inherently rare in the population (so that no sample of a realistic size would ever push their expected counts to 5) or where gathering additional data is not practically possible.

Ví dụ mẫu · Combining categories to fix a condition violation

A geneticist hypothesizes that eye color in a population follows Brown 45%, Blue 27%, Green 9%, Hazel 15%, Other 4%. In a sample of $n = 50$, the expected counts are Brown 22.5, Blue 13.5, Green 4.5, Hazel 7.5, Other 2.0 — Green and Other both fall below 5.
Remedy: combine Green and Other into a single "Green/Other" category. Its combined hypothesized proportion is $9\% + 4\% = 13\%$, giving a combined expected count of $0.13 \times 50 = 6.5 \geq 5$; all four remaining categories now satisfy the condition, and df drops from $5 - 1 = 4$ to $4 - 1 = 3$.
Observed counts were Brown 20, Blue 15, Green 3, Hazel 9, Other 3; after combining, Green/Other has $O = 3 + 3 = 6$, $E = 4.5 + 2.0 = 6.5$.

$$\chi^2 = \frac{(20-22.5)^2}{22.5} + \frac{(15-13.5)^2}{13.5} + \frac{(9-7.5)^2}{7.5} + \frac{(6-6.5)^2}{6.5} \approx 0.278 + 0.167 + 0.300 + 0.038 \approx 0.783.$$

With $df = 3$, the critical value at $\alpha = 0.05$ is 7.815. Since $0.783 < 7.815$, **fail to reject** H_0: the sample is consistent with the hypothesized (combined) eye-color distribution.

The same remedy applies to two-way tables. Suppose a 2×4 table testing independence between class year (Underclass/Upperclass) and favorite elective among four choices has one column — say,

"Foreign Language" — with very few students overall, giving expected counts below 5 in both rows of that column. The fix is identical in spirit: merge the "Foreign Language" column with a conceptually related column (for instance, another humanities elective) until every expected count in the merged table reaches 5, updating df from $(2-1)(4-1)=3$ before merging to $(2-1)(3-1)=2$ afterward, since the table now has one fewer column.

(!) Lỗi thường gặp: Chỉ gộp nhóm khi các nhóm có **ý nghĩa hợp lý để gộp chung** (ví dụ: hai màu mắt hiếm, hai mức thu nhập liền kề) — không được gộp tuỳ tiện chỉ để "sửa" cho expected ≥ 5 nếu việc gộp làm mất đi thông tin quan trọng của câu hỏi nghiên cứu. Sau khi gộp, phải nhớ **cập nhật lại** df theo số nhóm mới, và diễn giải kết luận theo đúng các nhóm đã gộp.

Formula reference for this unit.

Test statistic (all three tests): $\chi^2 = \sum \dfrac{(O-E)^2}{E}$.

Goodness of fit: $E_i = np_i, \quad df = k-1$.

Independence / homogeneity: $E_{ij} = \dfrac{\text{row total} \times \text{column total}}{n}, \quad df = (r-1)(c-1)$.

Conditions for all three tests: Random; 10% (if sampling without replacement); every expected count ≥ 5.

8.8 Practice problems

8.8.1 Goodness-of-fit practice

Bài tập luyện tập

A die is rolled 120 times: observed face counts are $18, 22, 19, 15, 25, 21$. Test H_0 : the die is fair at $\alpha = 0.05$.

(A) $\chi^2 = 3.0$, $df = 5$, critical $= 11.070$; fail to reject H_0

(B) $\chi^2 = 3.0$, $df = 6$, critical $= 12.592$; fail to reject H_0

(C) $\chi^2 = 6.0$, $df = 5$; reject H_0

(D) $\chi^2 = 3.0$, $df = 5$; reject H_0

Lời giải chi tiết

Expected each $= 120/6 = 20$. $\chi^2 = \dfrac{4+4+1+25+25+1}{20} = \dfrac{60}{20} = 3.0$. $df = 6-1 = 5$; critical value 11.070. Since $3.0 < 11.070$ (in fact 3.0 is even below the $\alpha = 0.10$ critical value of 9.236, so $P > 0.10$), **fail to reject** H_0 — consistent with a fair die. **(A)**.

Bài tập luyện tập

A retailer expects payment shares of Card 50%, Cash 30%, Mobile 20% among all transactions. In a random sample of 150 transactions, observed counts are Card 84, Cash 40, Mobile 26. Test H_0 at $\alpha = 0.05$.

(A) $\chi^2 \approx 2.17$, $df = 2$; fail to reject H_0

(B) $\chi^2 \approx 2.17$, $df = 3$; fail to reject H_0

(C) $\chi^2 \approx 4.17$, $df = 2$; reject H_0

(D) $\chi^2 \approx 2.17$, $df = 2$; reject H_0

Lời giải chi tiết

Expected: Card $= 0.50(150) = 75$, Cash $= 0.30(150) = 45$, Mobile $= 0.20(150) = 30$. $\chi^2 = \frac{9^2}{75} + \frac{(-5)^2}{45} + \frac{(-4)^2}{30} = 1.08 + 0.556 + 0.533 \approx 2.17$. $df = 3 - 1 = 2$; critical value at $\alpha = 0.05$ is 5.991, and at $\alpha = 0.10$ is 4.605; since 2.17 is below both, $P > 0.10$. **Fail to reject** H_0. **(A).**

Bài tập luyện tập

In a goodness-of-fit test with four categories, the individual components $(O - E)^2/E$ are $0.8, 3.6, 0.2, 1.4$. (a) What is χ^2? (b) With $df = 3$, is the result significant at $\alpha = 0.05$? (c) Which category contributed most to the test statistic?

Lời giải chi tiết

(a) $\chi^2 = 0.8 + 3.6 + 0.2 + 1.4 = 6.0$. (b) $df = 4 - 1 = 3$; critical value at $\alpha = 0.05$ is 7.815, and even at $\alpha = 0.10$ it is 6.251; since 6.0 is below both, $P > 0.10$. **Fail to reject** H_0. (c) The category with component 3.6 contributed the most — it alone accounts for $3.6/6.0 = 60\%$ of the total statistic.

Bài tập luyện tập

A sample of $n = 40$ is spread across five categories with hypothesized proportions $10\%, 15\%, 20\%, 25\%, 30\%$. Which of the following is true of the Large Counts condition for a chi-square goodness-of-fit test here?

(A) All expected counts are ≥ 5; condition met

(B) Violated, because the expected count for the 10% category is only 4, below 5

(C) Violated, because $n < 50$

(D) Violated, because $df < 4$

Lời giải chi tiết

Expected counts: $0.10(40) = 4$, $0.15(40) = 6$, $0.20(40) = 8$, $0.25(40) = 10$, $0.30(40) = 12$ (sum $= 40$, checks out). The first category has expected count $4 < 5$, so the Large Counts condition is violated; categories would need to be combined (e.g., merge the 10% and 15% categories) before proceeding. **(B).**

Bài tập luyện tập

Mendel's dihybrid-cross model predicts a phenotype ratio of $9{:}3{:}3{:}1$ (Round Yellow : Round Green : Wrinkled Yellow : Wrinkled Green) among $n = 160$ offspring. Observed counts are $88, 34, 28, 10$. Test whether the data are consistent with the $9{:}3{:}3{:}1$ ratio at $\alpha = 0.05$.

Lời giải chi tiết

State: H_0: the phenotype ratio is $9{:}3{:}3{:}1$, i.e. proportions $\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}$. H_a: the ratio is not $9{:}3{:}3{:}1$. $\alpha = 0.05$.
Plan: Chi-square goodness-of-fit test, $df = 4 - 1 = 3$. *Random:* assume offspring are a random sample from the cross. *10%:* 160 offspring are a small fraction of all possible offspring from the cross. *Expected counts:* $160(\frac{9}{16}) = 90$, $160(\frac{3}{16}) = 30$, $160(\frac{3}{16}) = 30$, $160(\frac{1}{16}) = 10$ — all ≥ 5, condition met.

Do: $\chi^2 = \frac{(-2)^2}{90} + \frac{4^2}{30} + \frac{(-2)^2}{30} + \frac{0^2}{10} = 0.044 + 0.533 + 0.133 + 0 = 0.711$.
Conclude: With $df = 3$, the critical value at $\alpha = 0.05$ is 7.815, and even the $\alpha = 0.10$ critical value is 6.251; since 0.711 is far below both, $P > 0.10$. **Fail to reject** H_0: the data are consistent with the $9{:}3{:}3{:}1$ genetic ratio.

Bài tập luyện tập

A campus dining hall claims that lunch orders split Vegetarian 20%, Chicken 35%, Beef 30%, Fish 15%. In a random sample of 200 orders, the observed counts are Vegetarian 34, Chicken 80, Beef 54, Fish 32. Test the dining hall's claim at $\alpha = 0.05$.

Lời giải chi tiết

State: H_0: the true order proportions match the claim ($20\%, 35\%, 30\%, 15\%$). H_a: at least one proportion differs. $\alpha = 0.05$.
Plan: Chi-square goodness-of-fit test, $df = 4 - 1 = 3$. *Random:* the 200 orders are a random sample. *10%:* 200 orders are less than 10% of a semester's worth of orders. *Expected:* $0.20(200) = 40$, $0.35(200) = 70$, $0.30(200) = 60$, $0.15(200) = 30$ — all ≥ 5, condition met.
Do: $\chi^2 = \frac{(-6)^2}{40} + \frac{10^2}{70} + \frac{(-6)^2}{60} + \frac{2^2}{30} = 0.900 + 1.429 + 0.600 + 0.133 \approx 3.06$.
Conclude: With $df = 3$, the critical value at $\alpha = 0.05$ is 7.815, and even at $\alpha = 0.10$ it is 6.251; since 3.06 is below both, $P > 0.10$. **Fail to reject** H_0: the sample is consistent with the dining hall's claimed order distribution.

Bài tập luyện tập

A city traffic department claims that car accidents are equally likely to occur on any of the seven days of the week. Records for a random sample of $n = 350$ accidents give the following counts by day: Mon 42, Tue 40, Wed 45, Thu 48, Fri 65, Sat 70, Sun 40. Test the department's claim at $\alpha = 0.05$.

Lời giải chi tiết

State: H_0: accidents are equally likely on each of the seven days ($p_i = \frac{1}{7}$ for each day). H_a: accidents are not equally likely across days. $\alpha = 0.05$.
Plan: Chi-square goodness-of-fit test, $df = 7 - 1 = 6$. *Random:* the 350 accidents are a random sample of all accident records. *10%:* 350 accidents are less than 10% of all accidents that could ever be recorded. *Expected counts:* $350/7 = 50$ for every day, all ≥ 5 — condition met.
Do:

$$\chi^2 = \frac{(-8)^2}{50} + \frac{(-10)^2}{50} + \frac{(-5)^2}{50} + \frac{(-2)^2}{50} + \frac{15^2}{50} + \frac{20^2}{50} + \frac{(-10)^2}{50}$$
$$= 1.28 + 2.00 + 0.50 + 0.08 + 4.50 + 8.00 + 2.00 = 18.36.$$

Conclude: With $df = 6$, the critical value at $\alpha = 0.05$ is 12.592, and at $\alpha = 0.01$ is 16.812. Since $18.36 > 16.812$, $P < 0.01$, so **reject** H_0: there is convincing evidence that accidents are not equally likely across the days of the week — Friday and Saturday, in particular, show noticeably more accidents than an equal-likelihood model predicts.

8.8.2 Independence practice

Bài tập luyện tập

A 3×4 table of counts is analyzed for independence and $\chi^2_{\text{calc}} = 15.2$. What is df, and what is the conclusion at $\alpha = 0.05$ (critical value 12.592)?

(A) $df = 7$; fail to reject

(B) $df = 6$; reject H_0, evidence of association

(C) $df = 12$; fail to reject

(D) $df = 6$; fail to reject

Lời giải chi tiết

$df = (r-1)(c-1) = (3-1)(4-1) = 6$. Since $15.2 > 12.592$, **reject** H_0: there is convincing evidence of association between the two variables. **(B)**.

Bài tập luyện tập

A survey classifies 180 people by exercise frequency (Low/Medium/High) and self-reported stress level (Low/High), a 3×2 table. The computed chi-square statistic is $\chi^2 = 8.44$. What is df, and what is the conclusion at $\alpha = 0.05$ (critical value at $df = 2$ is 5.991)?

(A) $df = 2$; reject H_0, evidence of association

(B) $df = 3$; fail to reject H_0

(C) $df = 6$; fail to reject H_0

(D) $df = 2$; fail to reject H_0

Lời giải chi tiết

$df = (3-1)(2-1) = 2$. Since $8.44 > 5.991$, **reject** H_0: there is convincing evidence of an association between exercise frequency and stress level. (Since 8.44 falls between the $\alpha = 0.025$ critical value of 7.378 and the $\alpha = 0.01$ critical value of 9.210, the exact P-value satisfies $0.01 < P < 0.025$.) **(A)**.

Bài tập luyện tập

A random sample of 200 shoppers is classified by preferred payment method (Card/Cash/Mobile) and generation (Gen Z/Millennial/Gen X or older). Test whether payment method and generation are associated, at $\alpha = 0.05$.

Generation	Card	Cash	Mobile	Total
Gen Z	35	15	20	70
Millennial	30	20	20	70
Gen X or older	25	15	20	60
Total	90	50	60	200

Lời giải chi tiết

State: H_0: payment method and generation are independent. H_a: payment method and generation are associated. $\alpha = 0.05$.
Plan: Chi-square test for independence, $df = (3-1)(3-1) = 4$. *Random:* the 200 shoppers are a random sample. *10%:* 200 shoppers are less than 10% of all shoppers. *Expected counts:*

Gen Z: $31.5, 17.5, 21.0$; Millennial: $31.5, 17.5, 21.0$; Gen X or older: $27.0, 15.0, 18.0$ — smallest is $15.0 \geq 5$, condition met.
Do:

$$\chi^2 = \frac{3.5^2}{31.5} + \frac{(-2.5)^2}{17.5} + \frac{(-1)^2}{21} + \frac{(-1.5)^2}{31.5} + \frac{2.5^2}{17.5} + \frac{(-1)^2}{21} + \frac{(-2)^2}{27} + \frac{0^2}{15} + \frac{2^2}{18}$$

$$\approx 0.389 + 0.357 + 0.048 + 0.071 + 0.357 + 0.048 + 0.148 + 0 + 0.222 \approx 1.64.$$

Conclude: With $df = 4$, the critical value at $\alpha = 0.05$ is 9.488, and even at $\alpha = 0.10$ it is 7.779; since 1.64 is well below both, $P > 0.10$. **Fail to reject** H_0: no convincing evidence that payment method and generation are associated.

Bài tập luyện tập

A survey of 400 respondents classified by education level (3 categories) and primary news platform (5 categories) forms a 3×5 table. The computed chi-square statistic is $\chi^2 = 22.1$. What is df, and what is the conclusion at $\alpha = 0.05$ (critical value at $df = 8$ is 15.507)?

(A) $df = 8$; reject H_0, evidence of association

(B) $df = 15$; fail to reject H_0

(C) $df = 8$; fail to reject H_0

(D) $df = 12$; reject H_0

Lời giải chi tiết

$df = (3-1)(5-1) = 8$. Since $22.1 > 15.507$, **reject** H_0: there is convincing evidence of an association between education level and primary news platform. **(A)**.

Bài tập luyện tập

A random sample of 150 college students is classified by class year (First-year/Upperclass) and car ownership (Yes/No). Test whether car ownership is associated with class year at $\alpha = 0.01$.

	Owns a car	No car	Total
First-year	18	57	75
Upperclass	42	33	75
Total	60	90	150

(A) $\chi^2 = 16.0$, $df = 1$; reject H_0

(B) $\chi^2 = 16.0$, $df = 2$; fail to reject H_0

(C) $\chi^2 = 8.0$, $df = 1$; reject H_0

(D) $\chi^2 = 16.0$, $df = 1$; fail to reject H_0

Lời giải chi tiết

Expected counts (row total$\times$col total$/150$): First-year–Owns $= 30$, First-year–No car $= 45$, Upperclass–Owns $= 30$, Upperclass–No car $= 45$ (all ≥ 5). $\chi^2 = \frac{(-12)^2}{30} + \frac{12^2}{45} + \frac{12^2}{30} + \frac{(-12)^2}{45} = 4.8 + 3.2 + 4.8 + 3.2 = 16.0$. $df = (2-1)(2-1) = 1$; critical value at $\alpha = 0.01$ is 6.635. Since $16.0 > 6.635$, **reject** H_0: convincing evidence that car ownership is associated with class year. **(A)**.

Bài tập luyện tập

A 2×3 table testing independence between smoking status (Smoker/Non-smoker) and preferred exercise type (Cardio/Strength/None) has expected counts $18, 24, 8, 22, 26, 12$. Are the conditions for the chi-square test met? Explain.

Lời giải chi tiết

Yes. All six expected counts $(18, 24, 8, 22, 26, 12)$ are at least 5 — the smallest is 8. As long as the data also came from a random sample (or randomized design) that is less than 10% of the population when sampled without replacement, all three conditions for the chi-square test for independence are satisfied.

Bài tập luyện tập

A two-proportion z-test on a 2×2 table gives $z = 2.83$. If the same data were analyzed as a chi-square test, what would be the value of χ^2 and its degrees of freedom?

(A) $\chi^2 \approx 8.01, df = 1$

(B) $\chi^2 \approx 2.83, df = 1$

(C) $\chi^2 \approx 8.01, df = 2$

(D) $\chi^2 \approx 16.02, df = 1$

Lời giải chi tiết

For a 2×2 table, $\chi^2 = z^2$, so $\chi^2 = 2.83^2 = 8.0089 \approx 8.01$; $df = (2-1)(2-1) = 1$. **(A)**.

Bài tập luyện tập

In the age-group/news-source example (Section 4), which single cell contributed the most to the total chi-square statistic, and what does that indicate?

Lời giải chi tiết

The 18–34/social-media cell has the largest component, 22.50, out of the total $\chi^2 \approx 69.87$ — about 32% of the entire test statistic on its own. This indicates that the youngest age group relies on social media as its primary news source far more than an independence model would predict ($O = 70$ vs. $E = 40$), and that this single cell is the main driver of the significant association found in that example.

8.8.3 Homogeneity practice

Bài tập luyện tập

A researcher gives three different weight-loss diets to three independent groups of patients and records, for each group, how many lost weight, stayed the same, or gained weight. Which chi-square test applies, and why?

(A) Goodness of fit, because there is one variable

(B) Independence, because there are two variables

(C) Homogeneity, because there are three separate groups (populations) compared on one categorical variable

(D) No chi-square test applies; use a t-test

Lời giải chi tiết

There are three separate populations/groups (the three diets), each measured on the same one categorical variable (weight change category) — this is a comparison of distributions across groups, i.e. a **chi-square test for homogeneity**. **(C)**.

Bài tập luyện tập

Random samples of 100 patients each are drawn from three hospitals and classified by treatment outcome (Improved/No change/Worsened), a 3×3 table. The computed chi-square statistic is $\chi^2 = 10.9$. What is df, and what is the conclusion at $\alpha = 0.05$ (critical value at $df = 4$ is 9.488)?

(A) $df = 4$; reject H_0, evidence outcome distributions differ

(C) $df = 4$; fail to reject H_0

(B) $df = 6$; reject H_0

(D) $df = 9$; reject H_0

Lời giải chi tiết

$df = (3-1)(3-1) = 4$. Since $10.9 > 9.488$, **reject** H_0: convincing evidence that the outcome distribution differs across the three hospitals. (Since 10.9 falls between the $\alpha = 0.05$ critical value of 9.488 and the $\alpha = 0.025$ critical value of 11.143, $0.025 < P < 0.05$.) **(A)**.

Bài tập luyện tập

Independent random samples of 100 students each are drawn from three high schools and classified by preferred elective category (STEM/Arts/Sports). Test whether the elective-preference distribution is the same across the three schools, at $\alpha = 0.05$.

School	STEM	Arts	Sports	Total
A	40	25	35	100
B	55	20	25	100
C	30	40	30	100
Total	125	85	90	300

Lời giải chi tiết

State: H_0: the distribution of elective preference is the same across the three schools. H_a: the distribution differs for at least one school. $\alpha = 0.05$.
Plan: Chi-square test for homogeneity, $df = (3-1)(3-1) = 4$. *Random:* independent random samples within each school. *10%:* each sample of 100 is a small fraction of its school's student body. *Expected counts:* $125/3 \approx 41.67$, $85/3 \approx 28.33$, $90/3 = 30.00$ for every school (row totals are all 100) — smallest is $28.33 \geq 5$, condition met.
Do:

$$\chi^2 = \frac{(-1.67)^2}{41.67} + \frac{(-3.33)^2}{28.33} + \frac{5^2}{30} + \frac{13.33^2}{41.67} + \frac{(-8.33)^2}{28.33} + \frac{(-5)^2}{30} + \frac{(-11.67)^2}{41.67} + \frac{11.67^2}{28.33} + \frac{0^2}{30}$$

$$\approx 0.067 + 0.392 + 0.833 + 4.267 + 2.451 + 0.833 + 3.267 + 4.804 + 0 \approx 16.91.$$

Conclude: With $df = 4$, the critical value at $\alpha = 0.05$ is 9.488 (and at $\alpha = 0.01$ is 13.277). Since $16.91 > 13.277$, $P < 0.01$, so **reject** H_0: convincing evidence that elective preference distribution differs across the three schools.

Bài tập luyện tập

Independent random samples of 60 trees each are taken from three separate forest plots (Plot 1, Plot 2, Plot 3), and each sampled tree is classified by health status (Healthy/Stressed/Dead). (a) Which chi-square test is appropriate for comparing the three plots? (b) What are the degrees of freedom?

Lời giải chi tiết

(a) There are three separate populations (the plots), each independently sampled and classified on the same one categorical variable (health status), so a **chi-square test for homogeneity** applies. (b) $df = (3-1)(3-1) = 4$.

Bài tập luyện tập

A study compares 4 independent groups on a 3-category outcome (a 4×3 table for homogeneity). If $\chi^2 = 14.9$, what is df, and what is the conclusion at $\alpha = 0.05$ (critical value at $df = 6$ is 12.592)?

(A) $df = 6$; reject H_0

(B) $df = 12$; fail to reject H_0

(C) $df = 6$; fail to reject H_0

(D) $df = 7$; reject H_0

Lời giải chi tiết

$df = (4-1)(3-1) = 6$. Since $14.9 > 12.592$, **reject** H_0: convincing evidence that the outcome distribution is not the same across the four groups. **(A)**.

Bài tập luyện tập

Independent random samples of 50 employees from two companies (X and Y) are classified by commute mode (Drive/Public transit/Walk-Bike). Test whether commute-mode distribution is the same for the two companies, at $\alpha = 0.05$.

Company	Drive	Transit	Walk/Bike	Total
X	30	12	8	50
Y	20	18	12	50
Total	50	30	20	100

(A) $\chi^2 = 4.0$, $df = 2$; fail to reject H_0

(B) $\chi^2 = 4.0$, $df = 3$; reject H_0

(C) $\chi^2 = 8.0$, $df = 2$; reject H_0

(D) $\chi^2 = 4.0$, $df = 1$; fail to reject H_0

Lời giải chi tiết

Expected counts (row total $= 50$ for each company): Drive $= 25$, Transit $= 15$, Walk/Bike $= 10$ for both companies (all ≥ 5). $\chi^2 = \frac{5^2}{25} + \frac{(-3)^2}{15} + \frac{(-2)^2}{10} + \frac{(-5)^2}{25} + \frac{3^2}{15} + \frac{2^2}{10} = 1.0 + 0.6 + 0.4 + 1.0 + 0.6 + 0.4 = 4.0$. $df = (2-1)(3-1) = 2$; critical value at $\alpha = 0.05$ is 5.991, and at $\alpha = 0.10$ is 4.605; since 4.0 is below both, $P > 0.10$. **Fail to reject** H_0. **(A)**.

Bài tập luyện tập

Independent random samples of 40 adults under age 40 and 40 adults aged 40 or older are classified by preferred beverage (Coffee/Tea/Neither). Test whether the beverage-preference distribution is the same for the two age groups, at $\alpha = 0.05$.

Age group	Coffee	Tea	Neither	Total
Under 40	25	10	5	40
40 or older	15	15	10	40
Total	40	25	15	80

(A) $\chi^2 \approx 5.17$, $df = 2$; fail to reject H_0

(B) $\chi^2 \approx 5.17$, $df = 3$; reject H_0

(C) $\chi^2 \approx 10.33$, $df = 2$; reject H_0

(D) $\chi^2 \approx 5.17$, $df = 1$; fail to reject H_0

Lời giải chi tiết

Expected counts (row totals equal 40): Coffee $= 40(40)/80 = 20$, Tea $= 40(25)/80 = 12.5$, Neither $= 40(15)/80 = 7.5$, the same for both age groups (all ≥ 5). $\chi^2 = \frac{5^2}{20} + \frac{(-2.5)^2}{12.5} + \frac{(-2.5)^2}{7.5} + \frac{(-5)^2}{20} + \frac{2.5^2}{12.5} + \frac{2.5^2}{7.5} = 1.25 + 0.5 + 0.833 + 1.25 + 0.5 + 0.833 \approx 5.17$. $df = (2-1)(3-1) = 2$; critical value at $\alpha = 0.05$ is 5.991. Since $5.17 < 5.991$, **fail to reject** H_0 — a close call, since 5.17 falls between the $\alpha = 0.10$ critical value (4.605) and the $\alpha = 0.05$ critical value (5.991), so $0.05 < P < 0.10$. **(A)**.

8.8.4 Choosing the correct test

Bài tập luyện tập

A single random sample of 500 registered voters is classified by political party and by their opinion on a ballot measure. Which chi-square test applies?

(A) Goodness of fit

(B) Independence

(C) Homogeneity

(D) No chi-square test applies

Lời giải chi tiết

One sample, two categorical variables measured on the same individuals, asking whether they are associated: **Independence**. **(B)**.

Bài tập luyện tập

A biologist samples 250 butterflies and classifies each by wing color, to check whether the color distribution matches a genetic model's predicted proportions. Which chi-square test applies?

(A) Goodness of fit

(B) Independence

(C) Homogeneity

(D) No chi-square test applies

Lời giải chi tiết

One sample, one categorical variable, compared to a distribution specified in advance (the genetic model): **Goodness of fit**. **(A)**.

Bài tập luyện tập

Independent random samples of 200 patients are drawn from each of four different hospitals and classified by satisfaction level (Low/Medium/High). Which chi-square test applies?

(A) Goodness of fit
(B) Independence
(C) Homogeneity
(D) No chi-square test applies

Lời giải chi tiết

Four separate populations (hospitals), each independently sampled, classified on the same one categorical variable, asking whether the distributions match across hospitals: **Homogeneity**. **(C)**.

8.8.5 Conditions and extra practice

Bài tập luyện tập

A goodness-of-fit test has hypothesized proportions $5\%, 10\%, 15\%, 30\%, 40\%$ and a sample size of $n = 45$. Which categories, if any, violate the expected-count condition?

(A) None; all expected counts are ≥ 5
(B) The 5% and 10% categories, with expected counts 2.25 and 4.5
(C) Only the 5% category
(D) All five categories

Lời giải chi tiết

Expected counts: $0.05(45) = 2.25, 0.10(45) = 4.5, 0.15(45) = 6.75, 0.30(45) = 13.5, 0.40(45) = 18$ (sum $= 45$). Both 2.25 and 4.5 are below 5, so those two categories violate the condition and should be combined with a neighboring category before testing. **(B)**.

Bài tập luyện tập

Explain why the chi-square statistic squares each deviation $(O-E)$ instead of simply summing the deviations directly. Address what would happen to $\sum(O-E)$ if it were used as the test statistic instead.

Lời giải chi tiết

Expected counts are built from the same totals (sample size n, and for independence/homogeneity, the same row and column totals) as the observed counts, so $\sum(O-E) = 0$ for *every* table — positive and negative deviations always cancel exactly, no matter how well or badly H_0 fits the data. Squaring each deviation makes every term non-negative, so genuine mismatches accumulate rather than canceling out, and it penalizes larger deviations more than proportionally (a deviation twice as large contributes four times as much). Dividing each squared deviation by E then standardizes it relative to the size of the expected count, so the same absolute miss counts for more when E is small than when E is large.

Bài tập luyện tập

As the degrees of freedom of a chi-square distribution increase, the shape of the curve:

(A) Becomes more strongly right-skewed

(B) Becomes more symmetric and bell-shaped, with its peak shifting to the right

(C) Becomes left-skewed

(D) Stays exactly the same regardless of df

Lời giải chi tiết

A chi-square variable with $df = k$ is the sum of k independent squared standard-normal variables. As k grows, the Central Limit Theorem pulls the sum's distribution toward a symmetric, approximately normal shape, and the mean of the distribution (which equals df) shifts rightward. **(B)**.

Bài tập luyện tập

A random sample of 120 students is classified by whether they eat breakfast regularly (Yes/No) and whether they pass a morning fitness test (Pass/Fail). Test whether breakfast consumption is associated with passing the fitness test, at $\alpha = 0.01$.

	Pass	Fail	Total
Breakfast: Yes	45	15	60
Breakfast: No	25	35	60
Total	70	50	120

Lời giải chi tiết

State: H_0: regular breakfast consumption and passing the fitness test are independent. H_a: they are associated. $\alpha = 0.01$.
Plan: Chi-square test for independence, $df = (2-1)(2-1) = 1$. *Random:* the 120 students are a random sample. *10%:* 120 students are less than 10% of all students. *Expected counts:* row total $= 60$ for both groups, so expected $= 35.0$ (Pass) and 25.0 (Fail) for both breakfast groups — all ≥ 5, condition met.
Do:

$$\chi^2 = \frac{10^2}{35} + \frac{(-10)^2}{25} + \frac{(-10)^2}{35} + \frac{10^2}{25} = 2.857 + 4.0 + 2.857 + 4.0 \approx 13.71.$$

Conclude: With $df = 1$, the critical value at $\alpha = 0.01$ is 6.635. Since $13.71 > 6.635$, **reject** H_0: convincing evidence that regular breakfast consumption is associated with passing the morning fitness test.

Bài tập luyện tập

A statistics software program reports $\chi^2 = 9.50$, $df = 3$, P-value $= 0.0234$ for a chi-square test. (a) Using $\alpha = 0.05$, what is the conclusion? (b) Show that this P-value is consistent with where $\chi^2 = 9.50$ falls in the critical-value table.

Lời giải chi tiết

(a) With $df = 3$, the critical value at $\alpha = 0.05$ is 7.815. Since $9.50 > 7.815$, **reject** H_0. (b) With $df = 3$, the critical values are 6.251 ($\alpha = 0.10$), 7.815 ($\alpha = 0.05$), 9.348 ($\alpha = 0.025$), and 11.345

($\alpha = 0.01$). Since $9.348 < 9.50 < 11.345$, the true P-value must lie between 0.01 and 0.025 — consistent with the software's reported value of $P = 0.0234$.

Inference for Quantitative Data: Slopes

Mục tiêu Unit: Điều kiện LINER cho hồi quy tuyến tính (đọc biểu đồ phần dư và Normal probability plot); kiểm định và khoảng tin cậy t (một đuôi và hai đuôi) cho hệ số góc β_1 của đường hồi quy tổng thể, $df = n - 2$; đọc và khai thác đầy đủ bảng kết quả hồi quy (computer regression output: b_0, b_1, SE_{b_1}, s, R-sq, T, P); liên hệ với kỹ thuật biến đổi dữ liệu (transformation) đã học ở Unit 2 khi điều kiện Linear không thỏa.

Unit 9 closes the course by returning to the least-squares regression line of Unit 2 and asking an inference question: does the linear pattern seen in *this* sample provide convincing evidence of a genuine linear relationship in the *population*, or could the observed slope simply be due to sampling variability? The mechanics are strikingly close to the one-sample t-procedures of Unit 7 — the same statistic-over-standard-error structure, the same t-distribution, the same four-step write-up — but applied to a new parameter, the population slope β_1, and gated by a new set of conditions, LINER. By the end of the unit, a full computer regression printout — a wall of numbers to an untrained eye — should read as fluently as a one-sample t-test summary: parameter, statistic, standard error, degrees of freedom, and conclusion. The unit also closes the loop back to Unit 2's least-squares line and Unit 3's study-design vocabulary, and forward to the kind of reasoning used in any applied field that fits a line to data — economics, medicine, environmental science, and beyond.

9.1 The four-step framework for slope inference

Every inference procedure in this course, from a one-proportion z-test in Unit 6 to a chi-square test in Unit 8, is organized the same way. Slope inference is no exception: the parameter changes to β_1, the standard error changes to SE_{b_1}, and the degrees of freedom change to $n - 2$, but the underlying logic — state a hypothesis or confidence level, plan a procedure and check conditions, do the calculation, and conclude in context — never changes.

The four-step framework, applied to slopes

- **State:** identify the parameter (β_1, the true slope of the population regression line), the hypotheses $H_0 : \beta_1 = 0$ vs. an appropriate H_a (for a significance test), or the confidence level $C\%$ (for an interval), and the significance level α if one is given.
- **Plan:** name the correct inference procedure (a t-test or t-interval for a slope) and verify the **LINER** conditions using a scatterplot, a residual plot, a Normal probability plot of residuals, and a description of how the data were produced.
- **Do:** if conditions are reasonably met, compute the test statistic $t = \dfrac{b_1 - 0}{SE_{b_1}}$ with $df = n - 2$ (for a test), or the interval $b_1 \pm t^* SE_{b_1}$ (for a confidence interval), reading b_1 and SE_{b_1} directly from computer output rather than computing them by hand.
- **Conclude:** state the decision — reject or fail to reject H_0, or report and interpret the interval — in a full sentence, in the context of the problem, in terms of β_1. Never write "accept H_0," and never claim certainty.

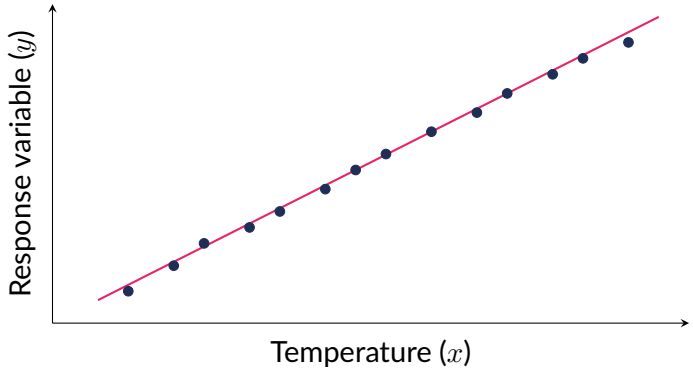
VN **GIẢI THÍCH TIẾNG VIỆT**

Bốn bước **State – Plan – Do – Conclude** áp dụng cho MỌI bài suy diễn thống kê trong chương trình, kể cả suy diễn về hệ số góc β_1: **State** nêu tham số cần suy diễn, giả thuyết (hoặc độ tin cậy mong muốn); **Plan** chọn đúng thủ tục (t-test hoặc t-interval cho slope) và kiểm tra đủ điều kiện **LINER**; **Do** thực hiện phép tính (t hoặc khoảng tin cậy) dựa trên b_1, SE_{b_1} đọc trực tiếp từ bảng *computer output*, KHÔNG tự tính tay; **Conclude** kết luận bằng câu văn đầy đủ, gắn với ngữ cảnh đề bài, không bao giờ viết "accept H_0" hay khẳng định chắc chắn tuyệt đối.

On the free-response portion of the exam, each of the four steps typically corresponds to a specific piece of credit that a reader looks for: a correctly identified parameter and hypotheses (State), a named procedure with conditions checked (Plan), correct mechanics (Do), and a contextual conclusion linked back to the P-value or interval (Conclude). Skipping the Plan step — jumping straight to a calculation — is one of the most common ways a response loses credit, even when the arithmetic itself is entirely correct; naming the procedure and at least gesturing at LINER is worth doing even under time pressure.

The parameter-statistic pattern below should look familiar from every inference unit so far — only the symbols and the standard-error formula change:

Population parameter	Sample statistic	Where it was introduced
p	$\hat{p}$	Unit 6, one proportion
μ	$\bar{x}$	Unit 7, one mean
β_1	b_1	Unit 9, slope of the population regression line



A strong, positive, linear scatterplot with no visible curve is the starting point for slope inference. The **Linear** and part of the **Random**/**Independent** conditions can already be judged from this plot and the study design; **Equal SD** and **Normal** residuals require the residual plot and Normal probability plot of Section 9.2.

A first four-step significance test

A biology class records the chirp rate (y, chirps per minute) of $n = 14$ crickets along with the ambient temperature in °F (x). Computer output for the regression of chirp rate on temperature:

Predictor	Coef	SE Coef
Constant	-70.5	18.9
Temp	3.75	0.62

Conditions have been checked and are satisfied. Test, at $\alpha = 0.05$, whether temperature is a useful linear predictor of chirp rate.

State: let β_1 be the true slope of the population regression line relating chirp rate to tempera-

ture. $H_0 : \beta_1 = 0$ (no linear relationship) vs. $H_a : \beta_1 \neq 0$; $\alpha = 0.05$.
Plan: a t-test for the slope. LINER conditions are given as satisfied, so $df = n-2 = 14-2 = 12$.
Do:

$$t = \frac{b_1 - 0}{SE_{b_1}} = \frac{3.75}{0.62} \approx 6.05.$$

The two-sided critical value at $\alpha = 0.05$ with $df = 12$ is $t^* = 2.179$.
Conclude: since $6.05 > 2.179$ (equivalently $P < 0.05$), **reject** H_0: there is convincing evidence of a linear relationship between temperature and chirp rate. Because $b_1 = 3.75 > 0$, the association is positive — warmer temperatures are linked to a higher chirp rate, on average. (This kind of relationship is well documented in entomology, sometimes called Dolbear's Law after an 1897 study of exactly this pattern.)

9.2 Conditions for regression inference

Before any t-test or t-interval for β_1 can be trusted, the data must plausibly have come from the population regression model

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where ε represents random scatter of individual y-values about the true population line. The sample regression line $\hat{y} = b_0 + b_1 x$ is only an *estimate* of this unknown population line, built from one particular sample; b_0 and b_1 are statistics that vary from sample to sample, just like $\bar{x}$ or $\hat{p}$ vary from sample to sample. Whether that variability behaves the way the t-procedures assume depends on five conditions, remembered by the acronym **LINER**.

The population regression model is $y = \beta_0 + \beta_1 x + \varepsilon$. Conditions (**LINER**): **L**inear relationship (scatterplot and residual plot show no curve); **I**ndependent observations (10% condition); **N**ormal residuals (especially important for small n); **E**qual standard deviation of residuals across all x (no fan/funnel shape); **R**andom sample or random assignment.

VN **GIẢI THÍCH TIẾNG VIỆT**
Ghi nhớ điều kiện hồi quy bằng **LINER**: Linear (thẳng), Independent (độc lập), Normal (phần dư gần chuẩn), Equal SD (phương sai phần dư đều), Random (ngẫu nhiên). Luôn kiểm tra **biểu đồ phần dư**: nếu có hình cong → vi phạm Linear; nếu độ trải phần dư tăng/giảm dần theo x (hình loa) → vi phạm Equal SD.

Each letter protects a different piece of the inference machinery. If **Linear** fails, the straight-line model itself is the wrong description of the relationship, so a slope b_1 is not even measuring something meaningful. If **Independent** or **Random** fail, the sampling-distribution theory behind SE_{b_1} no longer applies, and any P-value or confidence level is unreliable, regardless of how large t turns out to be. If **Normal** or **Equal SD** fail badly, the t-distribution is a poor model for how b_1 actually varies from sample to sample, especially when n is small — though, as with the one-sample t-procedures of Unit 7, larger sample sizes make the test more robust to mild skewness in the residuals. Because the five conditions protect different parts of the argument, a full LINER check should never be shortened to "conditions are met" without naming which evidence (which plot, which fact about the design) supports each letter.

Condition	How it is checked	Consequence if violated
Linear	Scatterplot and residual plot show no curve	The straight-line model misrepresents the relationship; b_1 does not summarize it meaningfully
Independent	10% condition; description of how the data were produced	Standard error formulas underestimate true variability; overly narrow intervals, misleadingly small P-values
Normal	Normal probability plot of residuals is roughly linear	The t-distribution is a poor model for b_1, especially with small n (less serious as n grows)
Equal SD	Residual plot shows roughly constant vertical spread (no fan)	SE_{b_1}, which assumes constant spread at every x, becomes inaccurate
Random	Random sample or random assignment described	Sampling-distribution theory does not apply; conclusions cannot be generalized beyond the data at hand

(!) Lỗi thường gặp: **Independent** và **Random** tuy đều liên quan đến cách thu thập dữ liệu nhưng hỏi hai câu khác nhau: **Random** hỏi "dữ liệu có được lấy mẫu ngẫu nhiên (hoặc xử lý có phân ngẫu nhiên) hay không" — nếu KHÔNG, kết luận không thể suy rộng ra tổng thể; **Independent** (10% condition) hỏi "mẫu có nhỏ hơn khoảng 10% tổng thể không" khi lấy mẫu không hoàn lại — nếu mẫu quá lớn so với tổng thể, các quan sát không còn gần như độc lập với nhau. Hai điều kiện này cần được nêu và kiểm tra RIÊNG BIỆT, không gộp chung thành một câu.

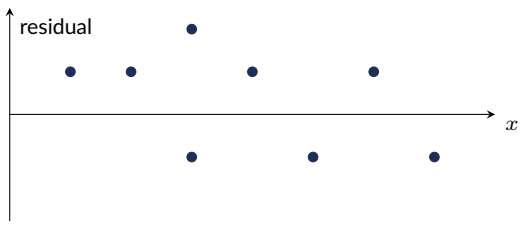
9.2.1 Checking LINER with residual plots and Normal probability plots

Two graphs do almost all the diagnostic work: a **residual plot** (residuals on the vertical axis, x or $\hat{y}$ on the horizontal axis) checks **Linear** and **Equal SD** simultaneously, while a **Normal probability plot** of the residuals checks **Normal**. **Independent** and **Random** are judged from how the data were collected, not from a plot.

Building a residual plot from raw data. A tutor records the number of study sessions attended (x) and quiz score out of 50 (y) for $n = 8$ students, and fits $\hat{y} = 20 + 3x$.

x	y (observed)	$\hat{y} = 20 + 3x$	residual $= y - \hat{y}$
1	24	23	1
2	27	26	1
3	28	29	-1
3	31	29	2
4	33	32	1
5	34	35	-1
6	39	38	1
7	40	41	-1

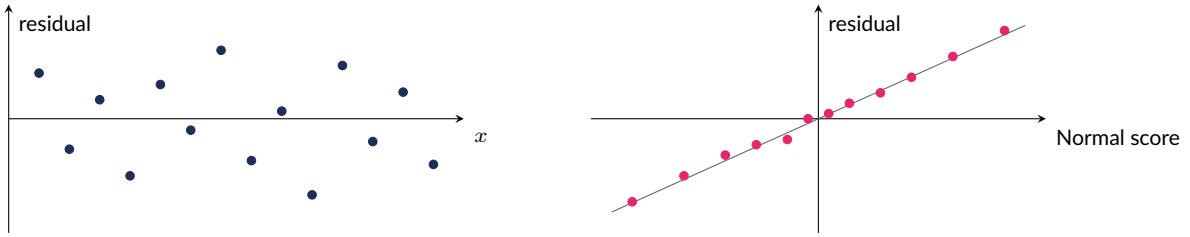
Each residual is computed as observed y minus predicted $\hat{y}$; plotting these eight $(x, \text{residual})$ points is exactly what produces a residual plot. Here the residuals bounce between about -1 and 2 with no upward or downward trend and no curve — the pattern that supports Linear and Equal SD. Software builds the same plot automatically from a much larger data set, but the underlying idea (subtract, then plot) never changes.



The residual plot built directly from the table above: no curve, no trend, roughly constant spread — exactly the pattern LINER requires.

Checking LINER: employees and monthly sales

A regional manager regresses monthly sales (y, in thousands of dollars) on the number of employees (x) for a random sample of $n = 18$ stores, representing less than 10% of all the company's stores. Two diagnostic plots are produced.



Left: residual plot — random scatter about zero, no curve, roughly constant spread. Right: Normal probability plot of the same residuals — points fall close to a straight diagonal line.

Determine whether inference for the slope is appropriate.

State: the LINER conditions must be checked before trusting any t-test or t-interval for β_1.

Plan: examine the residual plot for curvature (Linear) and changing spread (Equal SD); examine the Normal probability plot for departures from a straight line (Normal); use the description of the study for Independent and Random.

Do: the residual plot shows points scattered randomly above and below zero with no curved pattern and roughly the same vertical spread across all x — **Linear** and **Equal SD** both look satisfied. The Normal probability plot is close to a straight diagonal line, with no strong curvature at the ends — **Normal** is reasonably satisfied. The stores were randomly sampled and are fewer than 10% of all stores — **Independent** and **Random** are satisfied.

Conclude: all five LINER conditions are reasonably met, so it is appropriate to carry out t-based inference for the slope β_1 in this setting.

Checking LINER: a funnel-shaped residual plot

A different study regresses a car's stopping distance (y) on its speed (x) for $n = 24$ test runs. The residual plot shows residuals clustered tightly near zero for low predicted values, but fanning out much farther from zero as the predicted value increases; no curved pattern is visible, and a Normal probability plot of the residuals looks reasonably straight.

State/Plan: check LINER, focusing on the shape of the spread in the residual plot.

Do: no curvature in the residual plot $\Rightarrow$ **Linear** is satisfied; a fan/funnel shape (spread increasing with x) $\Rightarrow$ **Equal SD** is *violated*; a roughly straight Normal probability plot $\Rightarrow$ **Normal** is satisfied.

Conclude: because Equal SD fails, the standard SE_{b_1} formula (which assumes constant residual spread at every x) is not trustworthy here. Before performing t-based inference for the slope, a transformation of y (for example, a log or square-root transform, as in Unit 2) that stabilizes the spread should be tried.

(!) Lỗi thường gặp: Một hình cong (curved pattern) trong residual plot luôn có nghĩa vi phạm **Linear**, BẤT KỂ độ trải (spread) của các điểm có đều hay không. Ngược lại, hình loa (fan/funnel shape) — độ trải tăng hoặc giảm dần theo x — là vi phạm **Equal SD**, chứ KHÔNG phải Linear. Đây là hai lỗi khác nhau, dễ nhầm lẫn khi mô tả biểu đồ phần dư. Khi mô tả một residual plot

trên bài thi, luôn nói rõ CẢ HAI đặc điểm — hình dạng (có cong hay không) VÀ độ trải (đều hay loa) — vì mỗi đặc điểm ứng với một điều kiện LINER riêng biệt.

9.3 Inference for the slope β_1

If LINER holds, the sample slope b_1 behaves like any other statistic used earlier in the course: across repeated random samples of size n, the values of b_1 form an approximately Normal sampling distribution centered at the true population slope β_1, with some standard error SE_{b_1} that measures how much b_1 bounces around from sample to sample. Because the population standard deviation of the residuals is essentially never known, SE_{b_1} is estimated from the data, and the resulting standardized statistic follows a t-distribution rather than a standard Normal distribution — exactly as in the one-sample t-procedures for a mean. Intuitively, each parameter estimated from the same data set "uses up" one degree of freedom: a one-sample t-procedure for μ has $df = n - 1$ because only $\bar{x}$ is estimated, while slope inference has $df = n - 2$ because *both* b_0 and b_1 are estimated from the same n pairs of data before any residual can even be computed.

$H_0 : \beta_1 = 0$ (no linear relationship) is tested with

$$t = \frac{b_1 - 0}{SE_{b_1}}, \qquad df = n - 2,$$

where b_1 (sample slope) and SE_{b_1} are read from computer regression output. A CI for β_1: $b_1 \pm t^* SE_{b_1}$, $df = n - 2$.

VN **GIẢI THÍCH TIẾNG VIỆT**

Phân phối mẫu của b_1 (hệ số góc mẫu) có tâm tại β_1 (hệ số góc thật của tổng thể) khi các điều kiện LINER được thỏa mãn. SE_{b_1} đo độ dao động của b_1 giữa các mẫu ngẫu nhiên khác nhau cùng cỡ n — luôn LẤY TỪ bảng *computer output*, không tự tính tay trong hầu hết các bài. Thống kê kiểm định $t = \frac{b_1 - 0}{SE_{b_1}}$ dùng $df = n - 2$ (mất hai bậc tự do vì phải ước lượng cả b_0 lẫn b_1 từ cùng một bộ dữ liệu).

In this context, the P-value answers a specific question: if the true slope β_1 really were 0 (no linear relationship), how likely is it that a random sample of this size would produce a sample slope b_1 as far from 0 as — or farther than — the one actually observed? A small P-value means such an extreme sample slope would be unlikely under H_0, which counts as evidence against H_0; it is not, by itself, a statement about the size or practical importance of the relationship.

Hours studied and exam score

A regression of exam score on hours studied ($n = 12$) gives computer output:

Predictor	Coef	SE Coef
Constant	58.4	3.10
Hours	4.8	1.2

Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ at $\alpha = 0.05$, and give a 95% CI for β_1.

State: $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$, $\alpha = 0.05$, where β_1 is the true slope relating exam score to hours studied.

Plan: a t-test for the slope (LINER conditions assumed satisfied); $df = n - 2 = 12 - 2 = 10$.

Do:

$$t = \frac{4.8}{1.2} = 4.0, \qquad t^*_{df=10,\ \alpha=0.05} = 2.228.$$

Conclude: Since $4.0 > 2.228$, $P < 0.05$: **reject** H_0 — convincing evidence of a linear relationship between hours studied and exam score.

For the confidence interval (same conditions, $df = 10$, $t^* = 2.228$ for 95% confidence):

$$4.8 \pm 2.228(1.2) = 4.8 \pm 2.67 = (2.13,\ 7.47).$$

We are 95% confident that, on average, each additional hour of study is associated with an increase in exam score of between 2.13 and 7.47 points.

(!) Lỗi thường gặp: Không dùng công thức thủ công cho SE_{b_1} trừ khi đề yêu cầu — hầu hết bài AP cho sẵn bảng *computer output*; chỉ cần lấy đúng b_1 và SE_{b_1} từ bảng và nhớ $df = n - 2$ (khác với kiểm định trung bình dùng $df = n - 1$).

Is a specific slope value plausible? A confidence interval can also assess claims other than "$\beta_1 = 0$." Using the 95% CI from the example above, $(2.13,\ 7.47)$: is it plausible that the true slope is exactly 5 points per hour of study? Since 5 lies inside $(2.13,\ 7.47)$, this value is **plausible** — a two-sided test of $H_0 : \beta_1 = 5$ would fail to reject at $\alpha = 0.05$. Is a slope of exactly 1 point per hour plausible? Since 1 lies *outside* $(2.13,\ 7.47)$, it is **not** plausible — a test of $H_0 : \beta_1 = 1$ would reject at $\alpha = 0.05$. A confidence interval is really a whole collection of plausible hypothesized values, not just a check on whether 0 is one of them.

9.3.1 A one-sided significance test from computer output

Not every slope test is two-sided. When a researcher has a directional claim in mind before collecting data — "more of x is linked to *more* y," or "more of x is linked to *less* y" — the alternative hypothesis is one-sided, and the critical value comes from a single tail of the t-distribution rather than being split between two tails. Choosing the direction of H_a must happen *before* looking at the data; it should follow from the research question, not from whichever way the sample slope happened to point.

One-sided test: advertising hours and weekly sales

A retailer wants to know whether more hours of local TV advertising per week (x) are associated with *higher* weekly sales (y, in thousands of dollars), based on $n = 17$ weeks of data. LINER conditions have been checked and are satisfied. Computer output:

Predictor	Coef	SE Coef
Constant	12.4	5.30
Ad Hours	2.15	0.82

Test at $\alpha = 0.05$ whether there is convincing evidence of a *positive* linear relationship.

State: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 > 0$, $\alpha = 0.05$, where β_1 is the true slope relating weekly sales to advertising hours.

Plan: a t-test for the slope; conditions are given as satisfied; $df = 17 - 2 = 15$.

Do:

$$t = \frac{2.15}{0.82} \approx 2.62.$$

The one-sided critical value at $\alpha = 0.05$ with $df = 15$ is $t^* = 1.753$.

Conclude: since $2.62 > 1.753$ (so $P < 0.05$), **reject** H_0: there is convincing evidence that more advertising hours are associated with higher weekly sales, on average.

(!) Lỗi thường gặp: Với kiểm định một đuôi (one-sided), phải dùng đúng giá trị t^* ở cột diện tích một đuôi tương ứng α (không chia đôi α như kiểm định hai đuôi). Đồng thời hướng của H_a (> 0 hay < 0) phải khớp với dấu của b_1 quan sát được — nếu b_1 trái dấu với H_a, kết luận ngay lập tức là **fail to reject** H_0 mà không cần tính t.

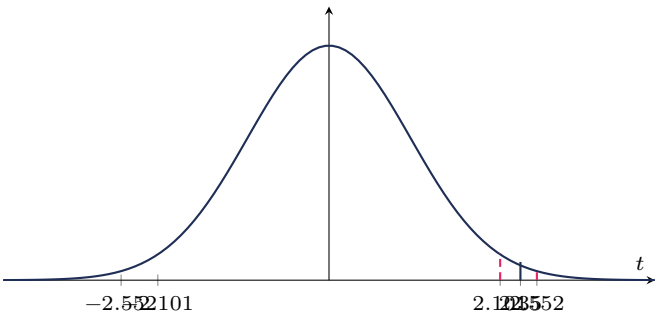
9.3.2 Bounding the P-value from the t-table

Many AP problems supply a printed t-table rather than an exact calculator P-value. The required skill is to locate $|t|$ between two adjacent critical values in the table and report the resulting bound on P, then compare that bound to α.

Bounding a P-value from the table

For a regression with $n = 20$ ($df = 18$), a slope test gives $t = 2.35$. Using the one-tail critical values for $df = 18$ (0.10, 0.05, 0.025, 0.01, 0.005 areas: 1.330, 1.734, 2.101, 2.552, 2.878), bound the P-value and state the two-sided conclusion at $\alpha = 0.05$.

Since $2.101 < 2.35 < 2.552$, the one-tail area satisfies $0.01 < P_{\text{one-tail}} < 0.025$; doubling for a two-sided test, $0.02 < P < 0.05$. Because this entire range lies below $\alpha = 0.05$, **reject** $H_0 : \beta_1 = 0$ even without an exact P-value from software.



The bell-shaped t-distribution ($df = 18$, shown for shape). The observed $t = 2.35$ lies between the tabulated critical values 2.101 (one-tail area 0.025) and 2.552 (one-tail area 0.01), so the one-tail P-value is bounded between 0.01 and 0.025.

A one-tailed bound. The same idea works for a one-sided test. Suppose $df = 25$ and a slope test gives $t = 1.95$, with H_a one-sided. The one-tail critical values for $df = 25$ are 1.316, 1.708, 2.060, 2.485, 2.787 (areas $0.10, 0.05, 0.025, 0.01, 0.005$). Since $t = 1.95$ exceeds 1.708 (the critical value for one-tail area 0.05) but not 2.060, the one-tail P-value satisfies $0.025 < P < 0.05$; because t is already beyond the 0.05 cutoff, $P < 0.05$, so a one-sided test at $\alpha = 0.05$ would reject H_0.

9.3.3 Recognizing calculator output (LinRegTTest)

Besides the Predictor/Coef/SE Coef table used by statistical software, the exam and most graphing calculators also present slope-inference results in a compact vertical format, typically labeled **LinRegTTest**. The pieces are identical to a table printout — only the layout changes. Both formats report the same underlying quantities because both come from the same least-squares calculation; a student who can read one fluently should expect to recognize the other on sight, without needing to relearn the procedure.

Calculator-style output: practice time and typing speed

A regression of typing speed (y, words per minute) on weeks of practice (x) for $n = 11$ students (LINER conditions satisfied) gives the following calculator output:

LinRegTTest

$\hat{y} = a + bx$	
$\beta \neq 0$ and $\rho \neq 0$	
t	$= 3.41$
df	$= 9$
a	$= 18.6$
b	$= 2.4$
s	$= 4.85$
r^2	$= 0.564$
r	$= 0.751$

Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ at $\alpha = 0.05$.

State: $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$, $\alpha = 0.05$, where β_1 is the true slope relating typing speed to weeks of practice.

Plan: a t-test for the slope; LINER given as satisfied; the calculator confirms $df = n - 2 = 11 - 2 = 9$.

Do: the calculator already reports $t = 3.41$. The two-sided critical value at $\alpha = 0.05$, $df = 9$, is $t^* = 2.262$.

Conclude: since $3.41 > 2.262$, **reject** H_0: convincing evidence of a linear relationship between weeks of practice and typing speed. The positive $r = 0.751$ matches the positive slope $b = 2.4$, and $r^2 = 0.564$: about 56.4% of the variability in typing speed is explained by weeks of practice.

9.4 Confidence interval for the slope β_1

A significance test only answers a yes/no question — is β_1 different from 0? — while a confidence interval estimates *how large* the true slope plausibly is, which is often the more useful piece of information in context (for example, a range of plausible dollar increases in revenue per dollar of advertising, rather than just "the effect is nonzero").

A confidence interval for the true slope β_1 has the same general form as every interval in this course — statistic $\pm$ (critical value)$\times$(SE of the statistic):

$$b_1 \pm t^* \, SE_{b_1}, \qquad df = n - 2,$$

where t^* is the two-sided critical value for the stated confidence level. The interval is interpreted exactly like any other CI: "We are $C\%$ confident that the interval from ___ to ___ captures the true population slope β_1."

VN **GIẢI THÍCH TIẾNG VIỆT**

Công thức khoảng tin cậy cho β_1 dùng đúng "khung sườn" của MỌI khoảng tin cậy trong chương trình: **statistic** $\pm$ **critical value** $\times$ **SE**. Với slope: $b_1 \pm t^* SE_{b_1}$, $df = n - 2$, và t^* luôn lấy từ cột **hai đuôi** (two-sided) dù bài kiểm định liên quan có thể một đuôi. Diễn giải luôn theo mẫu: "Chúng ta tin tưởng $C\%$ rằng khoảng từ ___ đến ___ chứa giá trị β_1 thật của tổng thể" — KHÔNG được nói về xác suất của β_1 (đó là một hằng số cố định, không phải biến ngẫu nhiên).

9.4.1 The link between the two-sided test and the confidence interval

From a confidence interval to a test decision

A 95% confidence interval for β_1, computed from a regression of plant height on fertilizer dose, is $(0.8,\ 3.4)$. Without repeating the t-test, what would a two-sided test of $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ conclude at $\alpha = 0.05$?

A $C\%$ confidence interval and a two-sided test at $\alpha = 1 - C$ always agree: if the hypothesized value (0) falls *outside* the interval, the corresponding test rejects H_0 at that α; if 0 falls *inside* the interval, the test fails to reject H_0. Here 0 is not contained in $(0.8,\ 3.4)$, so a two-sided test at $\alpha = 0.05$ would **reject** H_0: there is convincing evidence of a nonzero slope.

9.4.2 A 90% confidence interval from computer output

90% CI: daily exercise and resting heart rate

A researcher randomly samples $n = 22$ adults (fewer than 10% of the adult population) and regresses resting heart rate (y, in beats per minute) on average daily exercise time (x, in minutes). A scatterplot shows a linear, negative trend; the residual plot shows random scatter with roughly constant spread and no curve; a Normal probability plot of the residuals is approximately linear. Computer output:

Predictor	Coef	SE Coef
Constant	88.6	3.40
Exercise	-0.42	0.11

Construct and interpret a 90% confidence interval for β_1.

State: let β_1 be the true slope relating resting heart rate to daily exercise minutes; construct a 90% confidence interval for β_1.

Plan: a t-interval for the slope. LINER: **L**inear — the scatterplot is linear with no curve; **I**ndependent — $n = 22$ is less than 10% of all adults; **N**ormal — the Normal probability plot is roughly linear; **E**qual SD — the residual plot shows constant spread; **R**andom — the sample was randomly selected. All conditions are satisfied, so $df = 22 - 2 = 20$.

Do: the two-sided critical value for 90% confidence with $df = 20$ is $t^* = 1.725$.

$$-0.42 \pm 1.725(0.11) = -0.42 \pm 0.190 = (-0.610,\ -0.230).$$

Conclude: we are 90% confident that, on average, each additional minute of daily exercise is associated with a decrease in true mean resting heart rate of between 0.230 and 0.610 beats per minute.

Confidence level and interval width. As with every confidence interval in this course, a higher confidence level requires a larger critical value t^*, which widens the interval. Using the exercise/heart-rate data above ($b_1 = -0.42$, $SE_{b_1} = 0.11$, $df = 20$): a 90% CI uses $t^* = 1.725$ (margin of error 0.190), while a 99% CI would use $t^* = 2.845$ (margin of error $2.845(0.11) \approx 0.313$) — nearly 65% wider, in exchange for greater confidence that the interval captures β_1. There is no free lunch: for a fixed sample size, greater confidence always costs precision, whether the parameter is a mean, a proportion, or a slope. The other lever is sample size: because $SE_{b_1} = \dfrac{s}{s_x\sqrt{n-1}}$ shrinks as n grows, a larger study can achieve both high confidence *and* a narrow interval, which is exactly why researchers plan sample sizes in advance rather than accepting whatever data happen to be available.

9.4.3 When the interval contains zero

Occasionally a confidence interval for β_1 straddles 0 — for example, $(-1.4,\ 3.2)$. This does *not* mean β_1 is probably 0; it means the data are consistent with a wide range of slopes, including zero, negative values, and positive values. The correct conclusion is that these data do not provide convincing evidence of *either* a positive or a negative linear relationship — precisely the same conclusion a two-sided test would reach by failing to reject $H_0 : \beta_1 = 0$ at the matching significance level. A CI containing 0 is not evidence that no relationship exists, only that this particular data set could not distinguish the direction (or existence) of one with the desired confidence. For contrast, the exercise/heart-rate interval above, $(-0.610,\ -0.230)$, does *not* contain 0: every plausible value in that range is negative, so those data do provide convincing evidence of a genuinely negative relationship.

9.5 Reading full computer regression output

The AP exam almost always supplies regression results as computer output rather than raw data, so fluency in reading a full printout — not just picking out b_1 and SE_{b_1} — is essential.

Anatomy of a regression printout

For each predictor, a typical printout reports the coefficient estimate (**Coef**), its standard error (**SE Coef**), a t-statistic ($T =$ Coef/SE Coef, testing whether that coefficient is 0), and a P-value for that t-statistic. Below the coefficient table, **s** (the residual standard deviation, roughly the typical size of a residual, in the units of y) and **R-sq** ($= r^2$, the proportion of variability in y explained by the linear model on x) summarize the overall fit. The correlation $r = \pm\sqrt{r^2}$ always takes the same sign as b_1. Although the exam essentially always gives SE_{b_1} directly, it is worth knowing it is not arbitrary: $SE_{b_1} = \dfrac{s}{s_x\sqrt{n-1}}$, so a smaller residual spread s, a wider spread of x-values (s_x), or a larger sample size n all make SE_{b_1} smaller.

VN **GIẢI THÍCH TIẾNG VIỆT**

Đọc bảng *computer output* đầy đủ: hàng **Constant** cho b_0; hàng biến giải thích cho b_1 và SE_{b_1}; cột T chính là Coef/SE Coef (đã tính sẵn t-statistic, không cần tự chia); cột P là P-value (thường đã là hai đuôi). Bên dưới bảng, s là độ lệch chuẩn phần dư (các điểm dữ liệu thường lệch khỏi đường hồi quy khoảng s đơn vị của y), còn **R-sq** chính là r^2. Dấu của $r = \pm\sqrt{r^2}$ LUÔN trùng dấu với b_1: slope dương $\to r$ dương; slope âm $\to r$ âm.

Table printout	Calculator (LinRegTTest)	Meaning
Constant, Coef	a	intercept b_0
Predictor row, Coef	b	slope b_1
Predictor row, SE Coef	(computed as b/t)	SE_{b_1}
s	s	residual standard deviation
R-sq	r^2	coefficient of determination
Predictor row, T	t	test statistic for $H_0 : \beta_1 = 0$

Templates for interpreting regression output

Exam responses are scored on precise wording nearly as much as on correct arithmetic. Four templates cover almost every interpretation question in this unit:

- **Slope:** "For each additional [unit of x], the predicted [y, in context] [increases/decreases] by about [$|b_1|$] [units of y], on average."
- **Intercept b_0:** "When [x] $= 0$, the predicted [y] is [b_0]" — meaningful only if $x = 0$ is plausible

in context; otherwise the intercept is simply a mathematical anchor for the line, not a claim about real cases.

- r^2**:** "About [$r^2 \times 100$]% of the variability in [y, in context] is explained by the linear model on [x, in context]."
- **Test conclusion:** "Because P [$<$ or $\geq$] α, we [reject/fail to reject] H_0; the data [do/do not] provide convincing evidence of a linear relationship between [x] and [y] in the population."
- **Confidence interval:** "We are [C]% confident that the interval from [lower bound] to [upper bound] captures the true slope β_1 relating [y] to [x]."

VN **GIẢI THÍCH TIẾNG VIỆT**

Năm mẫu câu diễn giải trên nên thuộc lòng: (1) **slope** – "mỗi khi x tăng thêm 1 đơn vị, y dự đoán tăng/giảm trung bình khoảng $|b_1|$ đơn vị"; (2) **intercept** – chỉ có ý nghĩa thực tế nếu $x = 0$ hợp lý trong ngữ cảnh; (3) r^2 – luôn nói về % biến thiên của y được giải thích bởi mô hình tuyến tính trên x; (4) kết luận kiểm định – so sánh P với α rồi kết luận bằng ngữ cảnh; (5) khoảng tin cậy – luôn theo mẫu "tin tưởng C% rằng khoảng ... chứa β_1".

(!) Lỗi thường gặp: Diễn giải r^2 và diễn giải slope KHÔNG được đổi chỗ cho nhau: r^2 nói về **tỉ lệ biến thiên được giải thích** (một con số mô tả độ khớp mô hình), còn slope nói về **mức thay đổi trung bình của** y khi x tăng 1 đơn vị (một tốc độ thay đổi). Nói "mỗi đơn vị x tăng, y giải thích được r^2%" là sai — đó là hai đại lượng khác nhau hoàn toàn.

Applied to the engine-size example below: slope — "for each additional liter of engine size, predicted highway mpg decreases by about 4.35 mpg, on average"; r^2 — "about 64% of the variability in highway mpg is explained by the linear model on engine size"; test conclusion — "because $P = 0.000 < 0.05$, we reject H_0; the data provide convincing evidence of a linear relationship between engine size and fuel economy." Filling in each template with the actual numbers from a printout, rather than describing them vaguely, is what separates a full-credit response from a partial one.

A single **influential point** — often one with an unusually extreme x-value — can noticeably change b_1, SE_{b_1}, and R-sq even though it is only one observation among many; examining the scatterplot for such points, and noting whether removing them would change the conclusion, remains good practice even when a full printout is already provided.

Full output: engine size and fuel economy

A regression of highway fuel economy (y, in mpg) on engine size (x, in liters) for $n = 25$ cars gives:

Predictor	Coef	SE Coef	T	P
Constant	38.2	1.85	20.65	0.000
EngineSize	-4.35	0.68	-6.40	0.000

$s = 2.47$ mpg, R-sq $= 64.0\%$.

1. Write the least-squares regression equation.
2. Find and interpret r.
3. Is the slope significantly different from 0 at $\alpha = 0.05$ ($df = 23$)?
4. Predict the highway mpg for a car with a 2.5-liter engine.

1. $\hat{y} = 38.2 - 4.35x$.

2. R-sq $= 0.64$, so $r = \pm\sqrt{0.64} = \pm 0.8$; since $b_1 = -4.35 < 0$, $r = -0.8$ — a strong, negative linear association between engine size and fuel economy.
3. Check: $T = b_1/SE_{b_1} = -4.35/0.68 \approx -6.40$, matching the printed T-value; $df = 25 - 2 = 23$. This $|t|$ is far beyond every critical value in the t-table for nearby degrees of freedom, consistent with $P = 0.000 < 0.05$. **Reject** $H_0 : \beta_1 = 0$ — convincing evidence of a linear relationship between engine size and fuel economy.
4. $\hat{y} = 38.2 - 4.35(2.5) = 38.2 - 10.875 = 27.325 \approx 27.3$ mpg.

Full output without T/P: weekly mileage and race time

A coach regresses 5K race time (y, in minutes) on average weekly training mileage (x, in miles) for $n = 30$ runners; LINER conditions have been checked and are satisfied.

Predictor	Coef	SE Coef
Constant	26.4	1.10
Mileage	-0.085	0.028

(a) Test, at $\alpha = 0.01$, whether higher weekly mileage is associated with a *faster* (lower) race time. (b) Construct a 95% CI for β_1.

(a) State: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 < 0$, $\alpha = 0.01$; $df = 30 - 2 = 28$.
Plan/Do:

$$t = \frac{-0.085}{0.028} \approx -3.04.$$

The one-sided critical value at $\alpha = 0.01$ with $df = 28$ is $t^* = 2.467$.
Conclude: since $|-3.04| = 3.04 > 2.467$, **reject** H_0: convincing evidence that higher weekly mileage is associated with faster 5K times, on average.

(b) For a 95% CI with $df = 28$, the two-sided critical value is $t^* = 2.048$:

$$-0.085 \pm 2.048(0.028) = -0.085 \pm 0.0573 = (-0.142,\ -0.028).$$

We are 95% confident that each additional mile of average weekly training mileage is associated with a decrease in true mean 5K time of between 0.028 and 0.142 minutes.

Full output: a non-significant result

A researcher regresses weekly exercise hours (y) on weekly social-media hours (x) for $n = 20$ randomly selected college students (LINER conditions checked and satisfied):

Predictor	Coef	SE Coef	T	P
Constant	5.20	0.85	6.12	0.000
SocialMedia	-0.08	0.11	-0.73	0.478

$s = 1.35$ hours, R-sq $= 3.1\%$. Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ at $\alpha = 0.05$ ($df = 18$, $t^* = 2.101$), and comment on R-sq.

State: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, $\alpha = 0.05$; β_1 is the true slope relating exercise hours to social-media hours; $df = 20 - 2 = 18$.
Do: $T = -0.08/0.11 \approx -0.73$ (matches printed T); $|-0.73| = 0.73 < 2.101$.
Conclude: since $0.73 < 2.101$ (equivalently $P = 0.478 > 0.05$), **fail to reject** H_0: these data

do not provide convincing evidence of a linear relationship between social-media hours and exercise hours. This is consistent with R-sq $= 3.1\%$ — the linear model on social-media hours explains almost none of the variability in exercise hours, and $r = \pm\sqrt{0.031} \approx \pm 0.18$ is very close to 0.

9.6 Connecting back to transformations (Unit 2)

When LINER fails due to curvature

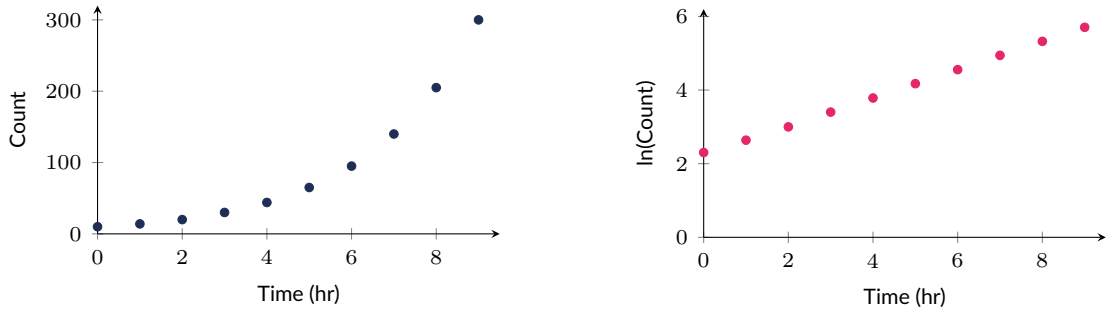
If a scatterplot (or residual plot) shows clear curvature, the Linear condition fails, and t-based inference for β_1 on the original variables is not valid — no matter how small SE_{b_1} looks. As in Unit 2, a **transformation** — commonly taking the logarithm of y (for exponential-type growth or decay) or a power/log-log transform of both variables (for power-law relationships) — can sometimes straighten the pattern. Slope inference is then carried out on the *transformed* variables, and conclusions are stated in terms of that transformed model.

VN **GIẢI THÍCH TIẾNG VIỆT**

Nếu biểu đồ phần dư có **hình cong** (vi phạm Linear), quay lại kỹ thuật **biến đổi dữ liệu** (transformation) đã học ở Unit 2: lấy log của y (khi dữ liệu tăng/giảm kiểu hàm mũ) hoặc log-log cả hai biến (khi dữ liệu theo hàm lũy thừa) để "làm thẳng" mối quan hệ, rồi mới áp dụng t-test/t-interval cho slope trên biến *đã biến đổi* — không áp dụng trực tiếp lên dữ liệu gốc còn cong.

Linearizing exponential growth

A biologist counts a bacterial colony (y) once every hour (x) for 10 hours. The scatterplot of y versus x curves sharply upward, so the Linear condition fails on the original scale. Taking the natural log of the counts and re-plotting $\ln(y)$ versus x produces a roughly straight scatterplot.



Left: raw colony count vs. time — clearly curved. Right: ln(count) vs. time — approximately linear, so slope inference can be carried out on this transformed model.

A regression of $\ln(y)$ on x gives a slope of $b_1 \approx 0.38$. Since the fitted model is $\ln(\hat{y}) = b_0 + 0.38x$, each additional hour multiplies the predicted count by $e^{0.38} \approx 1.46$ — an increase of about 46% per hour. Any t-test or t-interval for the slope should now be performed on this linear, transformed model rather than on the originally curved data.

The same idea extends to **power-law** relationships of the form $y = ax^p$: plotting $\ln(y)$ against $\ln(x)$ (a **log-log** transform, rather than logging y alone) straightens the pattern, and the slope of the fitted line on the log-log scale estimates the power p. Whichever transformation is used, the LINER conditions must be re-checked on the *transformed* scatterplot and residual plot before proceeding

with t-based inference — a straight $\ln(y)$-vs-x plot does not guarantee that the residuals from that transformed model are well-behaved, so the diagnostics are never skipped, only re-applied on the new scale.

With LINER established, the test statistic and confidence interval for β_1 computed, and a full computer printout readable end to end, the toolkit for quantitative-response regression is complete. Combined with the two-variable descriptive tools of Unit 2 and the study-design vocabulary of Unit 3, this unit closes the course's treatment of relationships between two variables — from first describing an association, to designing a study capable of investigating it, to finally deciding, with a stated level of confidence, whether it holds in the population.

9.7 Practice problems

(!) Lỗi thường gặp: Trước khi nộp một bài suy diễn về slope, tự kiểm tra nhanh: đã nêu đúng tham số β_1 trong ngữ cảnh (không chỉ viết chung chung "the slope") chưa? đã gọi tên đủ năm điều kiện LINER kèm bằng chứng cụ thể từ đề bài (không chỉ viết "conditions are met") chưa? đã dùng đúng $df = n - 2$ (không phải $n - 1$) chưa? câu kết luận có gắn với ngữ cảnh đề bài, so sánh P với α (hoặc t với t^*), và tránh các từ "accept H_0" hay "prove" không?

Summary of formulas — Unit 9: population model $y = \beta_0 + \beta_1 x + \varepsilon$; conditions **LINER**; test statistic $t = \dfrac{b_1 - 0}{SE_{b_1}}$, $df = n-2$; confidence interval $b_1 \pm t^* SE_{b_1}$, $df = n-2$; $r = \pm\sqrt{r^2}$ with the same sign as b_1; $SE_{b_1} = \dfrac{s}{s_x\sqrt{n-1}}$ (rarely computed by hand — read b_1 and SE_{b_1} directly from output). Whether the printout is a Predictor/Coef/SE Coef table or a calculator's LinRegTTest screen, the same five quantities (b_0, b_1, SE_{b_1}, s, r^2) drive every question in this unit.

Bài tập luyện tập

1. Regression output for $n = 15$: $b_1 = 2.3$, $SE_{b_1} = 0.9$. Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ at $\alpha = 0.05$ ($df = 13$, $t^* = 2.160$).

(A) $t \approx 2.56$; reject H_0

(B) $t \approx 0.39$; fail to reject H_0

(C) $t \approx 2.56$; fail to reject H_0

(D) $t \approx 1.44$; reject H_0

Lời giải chi tiết

$t = \dfrac{2.3}{0.9} \approx 2.56$. Since $2.56 > 2.160$, **reject** H_0 — convincing evidence of a nonzero slope. **(A)**.

Bài tập luyện tập

2. Using the same output ($b_1 = 2.3$, $SE_{b_1} = 0.9$, $df = 13$, $t^* = 2.160$), find the 95% confidence interval for β_1.

(A) $(0.36,\ 4.24)$

(B) $(1.40,\ 3.20)$

(C) $(-1.94,\ 4.24)$ (wrong sign somewhere)

(D) $(2.30,\ 2.30)$

Lời giải chi tiết

$ME = 2.160(0.9) \approx 1.944$. CI: $2.3 \pm 1.944 = (0.356,\ 4.244) \approx (0.36, 4.24)$. **(A)**.

Bài tập luyện tập

3. A regression gives $r^2 = 0.64$ and a *negative* slope. What is r?

(A) 0.64
(B) 0.8
(C) -0.8
(D) -0.64

Lời giải chi tiết

$r = \pm\sqrt{r^2} = \pm\sqrt{0.64} = \pm 0.8$; since the slope is negative, r must also be negative (they always share the same sign). $r = -0.8$. **(C)**.

Bài tập luyện tập

4. A residual plot for a regression of y on x shows a clear U-shape (curved pattern), and the spread of residuals is roughly constant across all x. Which LINER condition is violated, and which is satisfied?

Lời giải chi tiết

The curved pattern in the residual plot means the relationship between x and y is **not linear** — the **Linear** condition is violated, so a straight-line model (and slope inference based on it) is not appropriate; a transformation or different model should be considered. The roughly constant spread of residuals across x means the **Equal SD** condition *is* satisfied.

Bài tập luyện tập

5. A regression uses data from $n = 40$ subjects to estimate the population slope β_1. What are the degrees of freedom for the t-test/interval for the slope?

(A) 39
(B) 40
(C) 38
(D) 37

Lời giải chi tiết

For slope inference, $df = n - 2 = 40 - 2 = 38$ (two parameters, b_0 and b_1, are estimated from the same data). **(C)**.

Bài tập luyện tập

6. Which of the following, if true, would most likely *increase* SE_{b_1}, all else equal?

(A) A larger sample size n
(B) Data points lying closer to the regression line (smaller residual scatter)
(C) Data points lying farther from the regression line (larger residual scatter)
(D) A wider range of x-values in the sample

Lời giải chi tiết

Since $SE_{b_1} = \dfrac{s}{s_x\sqrt{n-1}}$, a *larger* residual standard deviation s (points farther from the line) makes SE_{b_1} larger. A larger n, a smaller s (points closer to the line), or a wider spread of x (s_x) would all make SE_{b_1} *smaller*, not larger. **(C)**.

Bài tập luyện tập

7. Regression output for $n = 20$: $b_1 = 1.8$, $SE_{b_1} = 0.65$. Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ at $\alpha = 0.05$ ($df = 18$, $t^* = 2.101$).

(A) $t \approx 2.77$; reject H_0
(B) $t \approx 1.17$; fail to reject H_0
(C) $t \approx 2.77$; fail to reject H_0
(D) $t \approx 0.36$; fail to reject H_0

Lời giải chi tiết

$t = \dfrac{1.8}{0.65} \approx 2.77$. Since $2.77 > 2.101$, **reject** H_0 — convincing evidence of a nonzero slope. **(A).**

Bài tập luyện tập

8. Regression output for $n = 10$: $b_1 = -0.9$, $SE_{b_1} = 0.4$. Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 < 0$ at $\alpha = 0.01$ ($df = 8$, one-tail $t^* = 2.896$).

(A) $t \approx -2.25$; reject H_0
(B) $t \approx -2.25$; fail to reject H_0
(C) $t \approx -3.60$; reject H_0
(D) $t \approx -0.44$; fail to reject H_0

Lời giải chi tiết

$t = \dfrac{-0.9}{0.4} = -2.25$. Since $|-2.25| = 2.25 < 2.896$, we **fail to reject** H_0: the data do not provide convincing evidence, at the strict $\alpha = 0.01$ level, of a negative linear relationship. **(B).**

Bài tập luyện tập

9. *(Four-step FRQ.)* A horticulturist regresses plant height after 6 weeks (y, cm) on fertilizer dose (x, grams) for $n = 13$ randomly selected plants (fewer than 10% of all plants of this type). LINER conditions have been checked and are satisfied. Output:

Predictor	Coef	SE Coef
Constant	8.10	1.60
Fertilizer	0.62	0.21

Test at $\alpha = 0.05$ whether there is a linear relationship between fertilizer dose and plant height.

Lời giải chi tiết

State: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, $\alpha = 0.05$, where β_1 is the true slope relating plant height to fertilizer dose.
Plan: t-test for the slope; conditions given as satisfied; $df = 13 - 2 = 11$.
Do: $t = \dfrac{0.62}{0.21} \approx 2.95$. Two-sided critical value, $df = 11$, $\alpha = 0.05$: $t^* = 2.201$.
Conclude: since $2.95 > 2.201$ ($P < 0.05$), **reject** H_0 — convincing evidence of a linear relationship between fertilizer dose and plant height.

Bài tập luyện tập

10. *(Four-step FRQ.)* A pediatric researcher regresses hours of sleep the night before an exam (y) on hours of screen time that evening (x) for $n = 10$ randomly chosen teenagers (fewer than 10% of all teenagers). LINER conditions are satisfied. Output:

Predictor	Coef	SE Coef
Constant	8.90	0.55
ScreenTime	-0.55	0.19

Test at $\alpha = 0.05$ whether there is a linear relationship between screen time and sleep hours.

Lời giải chi tiết

State: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, $\alpha = 0.05$, where β_1 is the true slope relating sleep hours to screen time.

Plan: t-test for the slope; conditions satisfied; $df = 10 - 2 = 8$.

Do: $t = \dfrac{-0.55}{0.19} \approx -2.89$. Two-sided critical value, $df = 8$, $\alpha = 0.05$: $t^* = 2.306$.

Conclude: since $|-2.89| = 2.89 > 2.306$ ($P < 0.05$), **reject** H_0 — convincing evidence of a (negative) linear relationship between screen time and sleep hours: more screen time is associated with less sleep, on average.

Bài tập luyện tập

11. *(Four-step FRQ.)* An education researcher wonders whether a student's seating distance from the front of the classroom (x, meters) is linearly related to test score (y), based on $n = 27$ randomly assigned seats (fewer than 10% of all possible seatings). LINER conditions are satisfied. Output:

Predictor	Coef	SE Coef
Constant	91.5	3.10
Distance	-1.2	0.55

Test at $\alpha = 0.05$ whether there is a linear relationship between seating distance and test score.

Lời giải chi tiết

State: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, $\alpha = 0.05$, where β_1 is the true slope relating test score to seating distance.

Plan: t-test for the slope; conditions satisfied; $df = 27 - 2 = 25$.

Do: $t = \dfrac{-1.2}{0.55} \approx -2.18$. Two-sided critical value, $df = 25$, $\alpha = 0.05$: $t^* = 2.060$.

Conclude: since $|-2.18| = 2.18 > 2.060$ ($P < 0.05$), **reject** H_0 — convincing evidence of a linear relationship between seating distance and test score (farther seats associated with lower scores, on average). Note this is an observational study, so no causal claim is justified — see Problem 26.

Bài tập luyện tập

12. Computer output for a slope test reports $T = 3.55$, $P = 0.002$. At $\alpha = 0.01$, what is the correct conclusion?

(A) Reject H_0, since $P < \alpha$
(B) Fail to reject H_0, since $P < \alpha$
(C) Reject H_0, since $T < \alpha$
(D) Fail to reject H_0, since $T > 3$

Lời giải chi tiết

Since $P = 0.002 < \alpha = 0.01$, the result is statistically significant: **reject** H_0. There is no need to compute t by hand when T and P are already printed. **(A)**.

Bài tập luyện tập

13. Regression output for $n = 16$: $b_1 = 4.1$, $SE_{b_1} = 1.15$. Find the 95% confidence interval for β_1 ($df = 14$, $t^* = 2.145$).

(A) $(1.63,\ 6.57)$
(B) $(2.95,\ 5.25)$
(C) $(-1.63,\ 6.57)$
(D) $(4.10,\ 4.10)$

Lời giải chi tiết

$ME = 2.145(1.15) \approx 2.467$. CI: $4.1 \pm 2.467 = (1.633,\ 6.567) \approx (1.63,\ 6.57)$. **(A)**.

Bài tập luyện tập

14. Regression output for $n = 32$: $b_1 = 0.28$, $SE_{b_1} = 0.09$. Find the 90% confidence interval for β_1 ($df = 30$, $t^* = 1.697$).

(A) $(0.13,\ 0.43)$
(B) $(0.19,\ 0.37)$
(C) $(0.28,\ 0.28)$
(D) $(-0.06,\ 0.62)$

Lời giải chi tiết

$ME = 1.697(0.09) \approx 0.153$. CI: $0.28 \pm 0.153 = (0.127,\ 0.433) \approx (0.13,\ 0.43)$. **(A)**.

Bài tập luyện tập

15. *(Four-step FRQ.)* A sleep researcher regresses reaction time change on a driving simulator (y, milliseconds) on hours of sleep the previous night (x) for $n = 22$ randomly selected drivers (fewer than 10% of all licensed drivers). Scatterplot, residual plot, and Normal probability plot all support LINER. Output:

Predictor	Coef	SE Coef
Constant	410	12.0
Sleep	-3.6	1.4

Construct and interpret a 95% confidence interval for β_1.

Lời giải chi tiết

State: let β_1 be the true slope relating reaction-time change to hours of sleep; construct a 95% CI.
Plan: t-interval for the slope; LINER satisfied (given); $df = 22 - 2 = 20$.

Do: two-sided critical value, $df = 20$, 95%: $t^* = 2.086$.

$$-3.6 \pm 2.086(1.4) = -3.6 \pm 2.9204 \approx (-6.52,\ -0.68).$$

Conclude: we are 95% confident that, on average, each additional hour of sleep is associated with a decrease in true mean reaction-time change of between 0.68 and 6.52 milliseconds.

Bài tập luyện tập

16. *(Four-step FRQ.)* A hardware store regresses weekly revenue from paint sales (y, dollars) on the number of DIY workshops held that week (x) for $n = 12$ randomly selected weeks (fewer than 10% of all weeks the store has operated). LINER conditions are satisfied. Output:

Predictor	Coef	SE Coef
Constant	210	40.0
Workshops	5.2	1.9

Construct and interpret a 95% confidence interval for β_1.

Lời giải chi tiết

State: let β_1 be the true slope relating weekly paint revenue to number of workshops; construct a 95% CI.
Plan: t-interval for the slope; LINER satisfied (given); $df = 12 - 2 = 10$.
Do: two-sided critical value, $df = 10$, 95%: $t^* = 2.228$.

$$5.2 \pm 2.228(1.9) = 5.2 \pm 4.2332 \approx (0.97,\ 9.43).$$

Conclude: we are 95% confident that, on average, each additional workshop held is associated with an increase in true mean weekly paint revenue of between $\$0.97$ and $\$9.43$.

Bài tập luyện tập

17. A 95% confidence interval for the slope relating advertising spend to revenue is $(1.2,\ 3.8)$ (revenue in thousands of dollars per thousand dollars spent). Which is the correct interpretation?

(A) 95% of the data points fall between a slope of 1.2 and 3.8

(B) There is a 95% probability that the true slope β_1 is between 1.2 and 3.8

(C) We are 95% confident that the interval from 1.2 to 3.8 captures the true population slope β_1

(D) 95% of all possible samples would give a slope in this exact interval

Lời giải chi tiết

A confidence interval is a statement about our confidence in the *method* capturing the fixed, unknown parameter β_1 — not a probability statement about β_1 itself (choice B), not a statement about individual data points (choice A), and not a claim that 95% of samples reproduce this exact interval (choice D — different samples give different intervals). **(C).**

Bài tập luyện tập

18. A different study regresses a car's stopping distance (y) on its speed (x). The residual plot shows residuals tightly clustered near zero for small predicted values but fanning out much wider for large predicted values, with no curved pattern. Which LINER condition is violated, and what should be done before running t-based inference?

Lời giải chi tiết

The fan/funnel shape (spread of residuals increasing with x, no curve) violates **Equal SD**, not Linear. The standard SE_{b_1} formula assumes constant residual spread, so it is unreliable here. A variance-stabilizing transformation of y (for example, a log or square-root transform) should be applied and the diagnostics re-checked before performing t-based inference for the slope.

Bài tập luyện tập

19. A Normal probability plot of residuals from a small study ($n = 9$) shows a pronounced S-shaped curve at both ends, though the middle of the plot is fairly straight. Is inference for the slope appropriate here?

Lời giải chi tiết

A pronounced curve at the ends of a Normal probability plot suggests the residuals may not be approximately Normal — the **Normal** condition is questionable. This concern is especially serious here because $n = 9$ is small; t-procedures are much less robust to non-Normal residuals with such a small sample. Inference for the slope should be treated with caution (or a transformation considered) rather than trusted at face value.

Bài tập luyện tập

20. Which single diagnostic is best for checking whether the residuals are approximately Normally distributed?

(A) The original scatterplot of y vs. x

(B) A residual plot (residuals vs. x or vs. predicted $\hat{y}$)

(C) A Normal probability plot of the residuals

(D) Knowing whether the data came from a random sample

Lời giải chi tiết

A residual plot (choice B) is designed to reveal curvature (Linear) and changing spread (Equal SD), not the shape of the residual distribution. A Normal probability plot (choice C) is specifically built to reveal departures from Normality — points falling close to a straight line support the Normal condition. **(C)**.

Bài tập luyện tập

21. A student wants to study the relationship between hours spent on social media (x) and self-reported stress level (y, on a 100-point scale), and recruits volunteers by posting a survey link in a single large school group chat. Explain what LINER condition this design threatens, and what limitation this places on any conclusions from a slope test or CI.

Lời giải chi tiết

This is a **convenience/voluntary-response sample**, not a random sample — the **Random** condition is violated, so the sampling-distribution theory behind the t-procedures does not strictly apply. Even if a very small P-value were found, the conclusion cannot be generalized beyond the students who happened to see and respond to the link; the **scope of inference** is limited to describing this particular volunteer group, not the broader population of students.

Bài tập luyện tập

22. A study's residual plot shows no curve and constant spread; its Normal probability plot of residuals is slightly curved at both ends but roughly linear in the middle; the $n = 45$ subjects were randomly selected and represent less than 10% of the population. Which LINER condition is most questionable, and how serious is the concern?

(A) Linear — serious concern

(B) Independent — serious concern

(C) Normal — mild concern given the sample size

(D) Random — serious concern

Lời giải chi tiết

Linear, Independent, and Random all look satisfied from the description. The Normal probability plot shows mild curvature at the ends, which flags the **Normal** condition, but with $n = 45$ (a reasonably large sample), the t-procedures are fairly robust to mild departures from Normality, so this is only a mild concern rather than a reason to abandon the analysis. **(C)**.

Bài tập luyện tập

23. *(Full-output FRQ.)* A realtor regresses home price (y, in thousands of dollars) on lot size (x, in thousands of square feet) for $n = 17$ randomly selected home sales (fewer than 10% of all sales in the area). LINER conditions have been checked and are satisfied. Output:

Predictor	Coef	SE Coef	T	P
Constant	145.2	12.4	11.71	0.000
LotSize	8.6	2.05	4.20	0.001

$s = 28.4$ (thousand dollars), R-sq $= 55.2\%$.

1. Write the least-squares regression equation.
2. Interpret the slope in context.
3. Find r.
4. Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ at $\alpha = 0.05$ ($df = 15$, $t^* = 2.131$).
5. Predict the price of a home on a $6{,}000$-square-foot lot ($x = 6$).

Lời giải chi tiết

1. $\hat{y} = 145.2 + 8.6x$.

2. For each additional $1{,}000$ square feet of lot size, predicted home price increases by about $\$8{,}600$, on average.

3. R-sq $= 0.552$, so $r = \pm\sqrt{0.552} \approx \pm 0.743$; since $b_1 > 0$, $r \approx 0.743$.

4. $t = 8.6/2.05 \approx 4.20$, matching the printed T. Since $4.20 > 2.131$ ($P = 0.001 < 0.05$), **reject** H_0 — convincing evidence of a linear relationship between lot size and home price.
5. $\hat{y} = 145.2 + 8.6(6) = 145.2 + 51.6 = 196.8$, i.e., about $\$196{,}800$.

Bài tập luyện tập

24. *(Full-output FRQ.)* A college advisor regresses first-semester GPA (y) on average weekly study hours (x) for $n = 20$ randomly selected students. LINER conditions have been checked and are satisfied. Output:

Predictor	Coef	SE Coef	T	P
Constant	2.10	0.18	11.67	0.000
StudyHrs	0.045	0.014	3.21	0.005

$s = 0.31$, R-sq $= 36.5\%$.

1. Write the least-squares regression equation, and find r.
2. A particular student studied $x = 10$ hours per week and earned an actual GPA of 2.80. Find and interpret the residual.
3. Construct a 95% confidence interval for β_1 ($df = 18$, $t^* = 2.101$).

Lời giải chi tiết

1. $\hat{y} = 2.10 + 0.045x$. R-sq $= 0.365$, so $r = \pm\sqrt{0.365} \approx \pm 0.604$; since $b_1 > 0$, $r \approx 0.604$.
2. Predicted GPA: $\hat{y} = 2.10 + 0.045(10) = 2.10 + 0.45 = 2.55$. Residual = observed − predicted $= 2.80 - 2.55 = 0.25$. The model **underpredicted** this student's GPA by 0.25 points.
3. $ME = 2.101(0.014) \approx 0.0294$. CI: $0.045 \pm 0.0294 = (0.0156,\ 0.0744)$. We are 95% confident that, on average, each additional weekly study hour is associated with an increase in true mean GPA of between 0.016 and 0.074 points.

Bài tập luyện tập

25. *(Full-output FRQ.)* A marketing analyst regresses monthly revenue (y, in thousands of dollars) on advertising budget (x, in thousands of dollars) for $n = 13$ randomly selected months. LINER conditions have been checked and are satisfied. Output:

Predictor	Coef	SE Coef	T	P
Constant	52.0	6.8	7.65	0.000
AdBudget	3.15	0.95	3.32	0.007

$s = 9.6$ (thousand dollars), R-sq $= 49.9\%$.

1. Interpret $s = 9.6$ in context.
2. Is the slope convincingly different from 0 at $\alpha = 0.01$ ($df = 11$, $t^* = 3.106$)?
3. Does R-sq $= 49.9\%$ mean advertising budget is not a useful predictor? Explain.

Lời giải chi tiết

1. The actual monthly revenue typically differs from the value predicted by the regression line by about $\$9{,}600$.

2. $t = 3.15/0.95 \approx 3.32$ (matches printed T). Since $3.32 > 3.106$ ($P = 0.007 < 0.01$), **reject** H_0 — the slope is convincingly different from 0 even at this strict level.

3. No. Statistical significance ($P = 0.007$) and the strength of fit (R-sq) answer different questions: the slope is convincingly nonzero, but advertising budget alone explains only about half (49.9%) of the variability in monthly revenue — other factors (season, competitor pricing, staffing, etc.) likely also matter. A significant slope does not require a high R-sq.

Bài tập luyện tập

26. In an observational study of $n = 200$ cities, a regression of heart-disease rate (y) on average commute time (x) gives $t = 5.8$ and $P < 0.001$. A newspaper headline claims, "Long commutes cause heart disease." Explain the flaw in this claim.

Lời giải chi tiết

This is an **observational study** — no treatments were randomly assigned to cities. A highly significant slope only shows a strong linear **association** between commute time and heart-disease rate, not **causation**. **Confounding variables** (for example, income, urban density, occupational stress, or exercise habits) could plausibly affect both commute time and heart-disease risk, producing the observed association without either variable directly causing the other. Only a well-designed, randomized **experiment** — which is not feasible here, since commute time cannot ethically or practically be randomly assigned to people — could support a causal claim.

Bài tập luyện tập

27. A residual plot for a regression of y on x shows a clear downward-then-upward curve (like the letter U), with roughly constant vertical spread. What should be done next?

(A) Report the slope and P-value as usual — the pattern does not matter

(B) Increase the sample size to fix the problem

(C) Recognize that the Linear condition fails and consider a transformation or a different (curved) model

(D) Switch to a chi-square test

Lời giải chi tiết

A curved residual pattern always signals a violation of the **Linear** condition, regardless of how large or constant the spread is, and regardless of sample size — more data would not fix a fundamentally wrong model shape. The appropriate response is to consider a transformation (Unit 2) or a nonlinear model before trusting any slope inference. Chi-square procedures are for categorical, not quantitative, data. **(C)**.

Bài tập luyện tập

28. A calculator's LinRegTTest screen for a regression of $n = 14$ observations reports $t = 1.95$, $df = 12$. At $\alpha = 0.05$ (two-sided, $t^* = 2.179$), what is the correct conclusion?

(A) Reject H_0, since $t > 0$

(B) Fail to reject H_0, since $1.95 < 2.179$

(C) Reject H_0, since $df = 12$ is large enough

(D) Fail to reject H_0, since $t < df$

Lời giải chi tiết

Compare $|t|$ to t^*: $1.95 < 2.179$, so **fail to reject** H_0 — the data do not provide convincing evidence of a nonzero slope. Reasoning about the sign of t (choice A) or the size of df alone (choices C, D) does not determine significance. **(B)**.

Bài tập luyện tập

29. *(Four-step FRQ.)* A gardener uses a calculator to regress tomato yield (y, kg) on hours of sunlight per day (x) for $n = 16$ randomly selected plants (fewer than 10% of all such plants). LINER conditions are satisfied. The calculator's LinRegTTest reports $b = 0.85$, $t = 3.10$, $df = 14$. Test at $\alpha = 0.05$ whether sunlight hours are linearly related to yield.

Lời giải chi tiết

State: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, $\alpha = 0.05$, where β_1 is the true slope relating tomato yield to sunlight hours.

Plan: t-test for the slope; LINER satisfied; $df = 16 - 2 = 14$, matching the calculator.

Do: the calculator reports $t = 3.10$ directly. Two-sided critical value, $df = 14$, $\alpha = 0.05$: $t^* = 2.145$.

Conclude: since $3.10 > 2.145$ ($P < 0.05$), **reject** H_0 — convincing evidence of a linear relationship between sunlight hours and tomato yield. Since $b = 0.85 > 0$, more sunlight is associated with higher yield, on average.

Bài tập luyện tập

30. A statistical-software printout reports "Coef" and "SE Coef" for each predictor, while a calculator's LinRegTTest reports a, b, t, and s. Which correctly matches the calculator notation to the table notation?

(A) $a =$ Coef for Constant; $b =$ Coef for the predictor

(B) $a =$ Coef for the predictor; $b =$ Coef for Constant

(C) $a = SE$ Coef; $b = T$

(D) $a = r$; $b = r^2$

Lời giải chi tiết

In $\hat{y} = a + bx$, a is the intercept, matching the **Constant** row's Coef in a table printout, and b is the slope, matching the predictor row's Coef. Likewise, the calculator's s corresponds to the printout's residual standard error, and its t corresponds to the predictor row's T value used in the significance test. **(A)**.

Ôn tập nhanh trước ngày thi:
Toàn cảnh 9 Unit:

1. **Unit 1 – Exploring One-Variable Data:** shape–center–spread–outlier; mean/median, SD/IQR; z-score; phân phối chuẩn và quy tắc 68–95–99.7.
2. **Unit 2 – Exploring Two-Variable Data:** scatterplot, hệ số tương quan r; đường hồi quy bình phương nhỏ nhất (LSRL) $\hat{y} = a + bx$; phần dư (residual); r^2; biến đổi dữ liệu (transformation) khi mối quan hệ không tuyến tính.
3. **Unit 3 – Collecting Data:** các phương pháp chọn mẫu (SRS, phân tầng, cụm, hệ thống); các loại sai lệch (bias); nguyên tắc thiết kế thí nghiệm; phạm vi suy luận (scope of inference) – chỉ thí nghiệm ngẫu nhiên hoá tốt mới cho phép kết luận nhân quả.
4. **Unit 4 – Probability, Random Variables, and Probability Distributions:** quy tắc xác suất; biến ngẫu nhiên rời rạc; phân phối nhị thức (binomial) và hình học (geometric); giá trị kỳ vọng và phương sai.
5. **Unit 5 – Sampling Distributions:** phân phối mẫu của $\bar{x}$ và $\hat{p}$; định lý giới hạn trung tâm (CLT); điều kiện Random – 10% – Large Counts/Normal.
6. **Unit 6 – Inference for Categorical Data: Proportions:** kiểm định và khoảng tin cậy z cho một và hai tỉ lệ.
7. **Unit 7 – Inference for Quantitative Data: Means:** kiểm định và khoảng tin cậy t cho một và hai trung bình (độc lập và bắt cặp – paired).
8. **Unit 8 – Inference for Categorical Data: Chi-Square:** khớp tốt (goodness of fit), độc lập (independence), đồng nhất (homogeneity); luôn dùng số đếm, expected ≥ 5.
9. **Unit 9 – Inference for Quantitative Data: Slopes:** điều kiện LINER; kiểm định và khoảng tin cậy t cho β_1, $df = n - 2$; đọc bảng *computer output* đầy đủ.

Tham số	Thống kê mẫu	SE (ước lượng)	Phân phối
p	$\hat{p}$	$\sqrt{\hat{p}(1-\hat{p})/n}$	z
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	z
μ	$\bar{x}$	$s/\sqrt{n}$	t, $df = n-1$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{s_1^2/n_1 + s_2^2/n_2}$	t, $df \approx \min(n_1, n_2) - 1$
β_1	b_1	SE_{b_1} (đọc từ output)	t, $df = n-2$

Ghi nhớ chung khi làm bài: (1) luôn nêu đủ shape–center–spread–outlier trong ngữ cảnh; (2) s_x chia cho $n-1$, còn $\sigma_{\bar{x}} = \sigma/\sqrt{n}$; (3) kiểm tra ĐỦ điều kiện (Random – 10% – Large Counts/Normal, hoặc LINER với hồi quy) trước mọi CI/test; (4) statistic $\pm$ (critical value)$\times$(SD của statistic) là công thức gốc của MỌI khoảng tin cậy; (5) t: $df = n-1$ (một mẫu), $df = n-2$ (hồi quy); (6) χ^2: luôn dùng số đếm, expected ≥ 5; (7) luôn viết "reject/fail to reject H_0", không bao giờ "accept H_0" hay nói xác suất về một tham số cố định.
Ôn lại đầy đủ công thức, điều kiện áp dụng, và cách diễn giải kết quả theo đúng ngữ cảnh của từng bài toán trước khi bước vào phòng thi.

Đồng hành cùng 8020 English

Cảm ơn bạn đã học cùng bộ giáo trình quốc tế 8020. **Điểm thật · Thầy thật · Kết quả thật** — nhiều học viên chỉ cần 1–2 khoá để đạt kết quả mong muốn.

Chương trình đào tạo tại 8020 English

IELTS Academic/General · **SAT** · PTE · Duolingo · TOEFL | **AP** (College Board) · **IB** Diploma · **Cambridge** IGCSE / A-Level | săn học bổng du học.

Vì sao chọn 8020 English?

- **100% giáo viên** đạt tối thiểu 8.0 IELTS hoặc 1500 SAT — đội ngũ Giáo sư · Tiến sĩ · Thạc sĩ tốt nghiệp trường top đầu thế giới (Carnegie Mellon — #1 Hoa Kỳ về Khoa học Máy tính).
- **Kèm 1–1** mọi độ tuổi; lớp sĩ số nhỏ 2–5 bạn (đa số dưới 10 bạn/lớp).
- Lộ trình cá nhân hoá, theo sát tiến độ từng học viên; lớp OFFLINE tại TP.HCM & ONLINE toàn quốc.

Bảng vàng học viên

AP 5/5 — Calculus AB/BC	**AP** 4–5 — Biology, Chemistry	**IB** 90/100 — Economics
SAT 1590 / 1570 / 1550	**IELTS** 8.0–9.0	**Duolingo** 110 · **PTE** 90

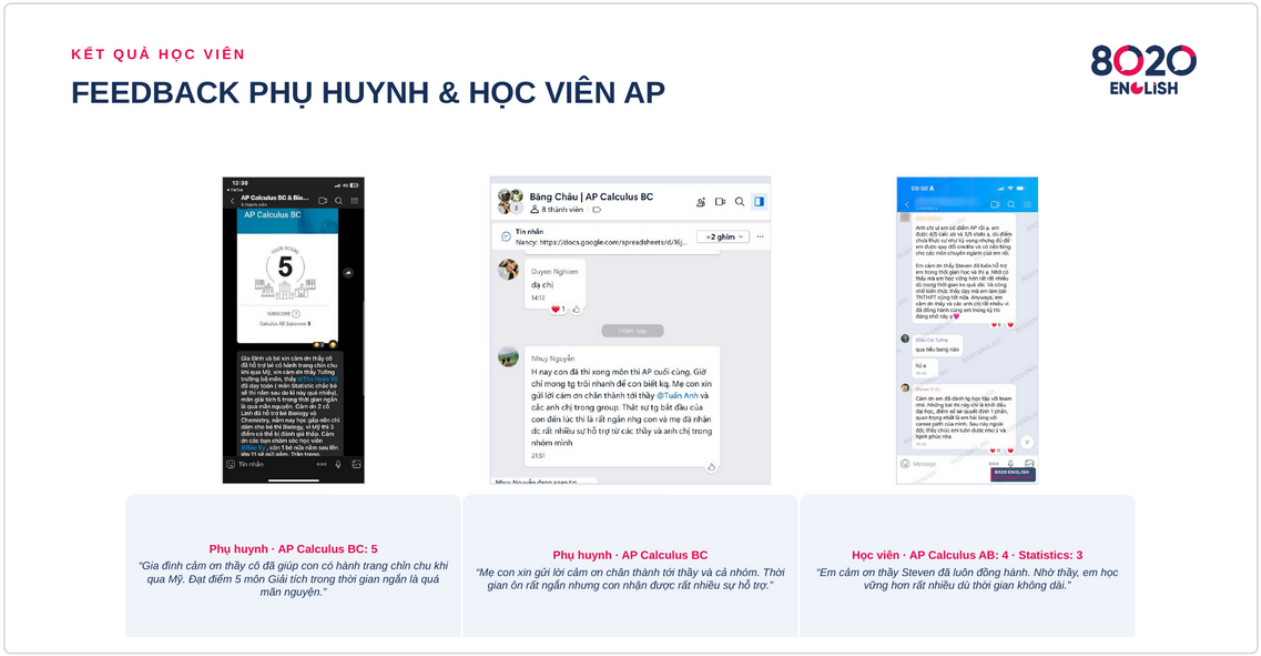
Cảm nhận học viên & phụ huynh

> *"Bạn đã cải thiện được tư duy Writing — kỹ năng từng là nỗi sợ lớn nhất — và cuối cùng đã đậu vào trường top như mong muốn. Thái độ học tập cực kỳ nghiêm túc; ngay cả khi nằm viện vẫn cố gắng học đều đặn."*
>
> — Phan Thị Thanh Hằng · IELTS Overall 8.5

> *"Cần tất cả kỹ năng trên 7.5 để xét vào trường top 1 Anh Quốc về Y học (chương trình UKFPO của NHS) — và đã đạt mục tiêu sau khi học Writing 1–1."*
>
> — Trần Hoàng Bảo Ngọc · IELTS Overall 8.0 (Writing 6.0 → 7.5)

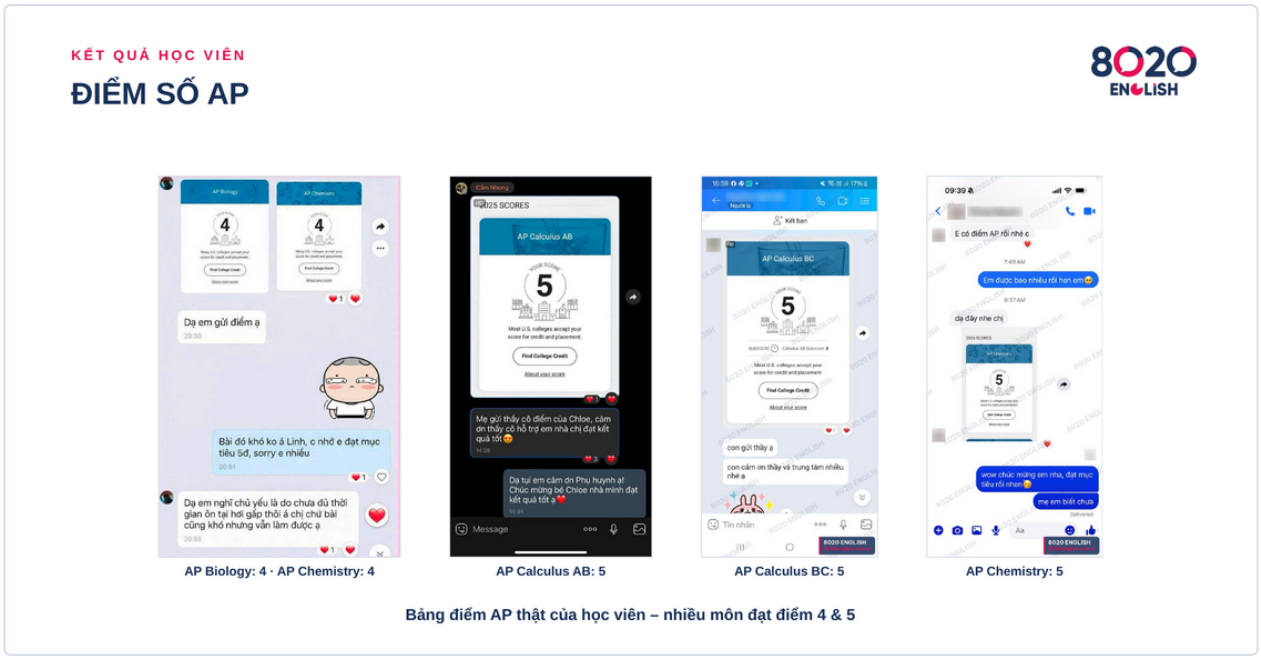
> *"Gia đình và bé xin cảm ơn thầy cô đã giúp con có hành trang chỉn chu khi qua Mỹ. Đạt điểm 5 môn Giải tích trong thời gian ngắn là quá mãn nguyện."*
>
> — Phụ huynh · AP Calculus BC 5/5

> *"Em cảm ơn thầy đã luôn đồng hành. Nhờ thầy, em học vững hơn rất nhiều dù thời gian ôn không dài."*
>
> — Học viên · AP Calculus AB 4

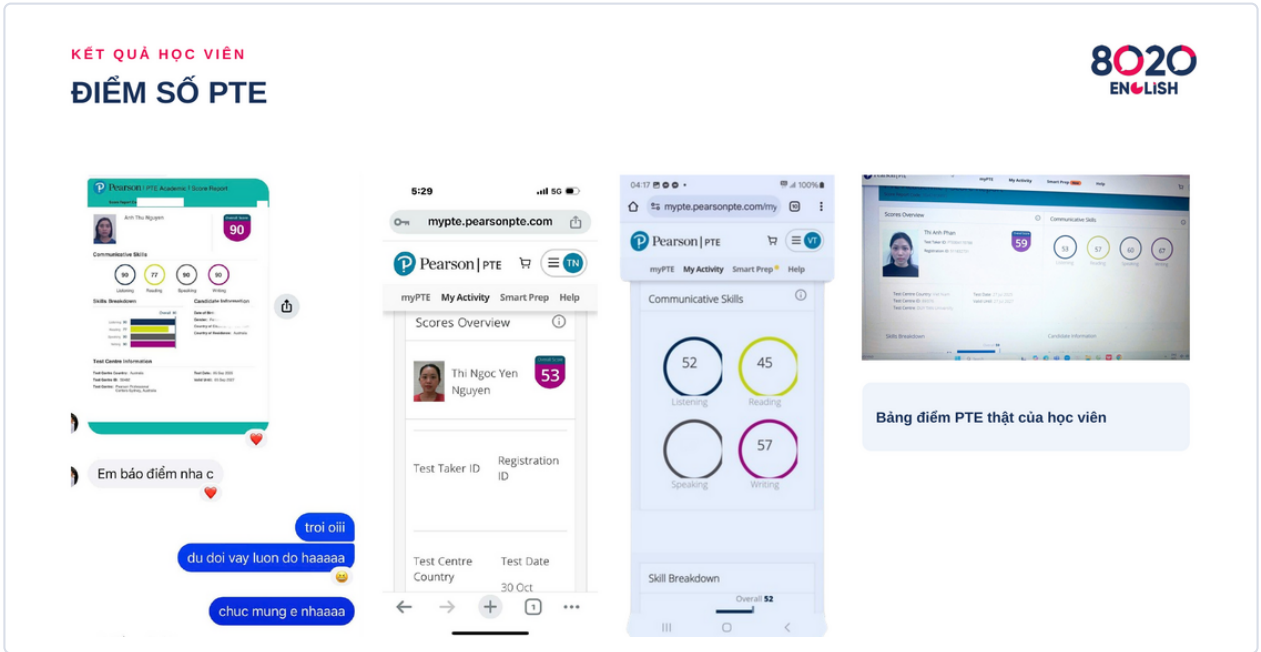


Feedback phụ huynh & học viên AP — nhiều môn đạt điểm 4 & 5

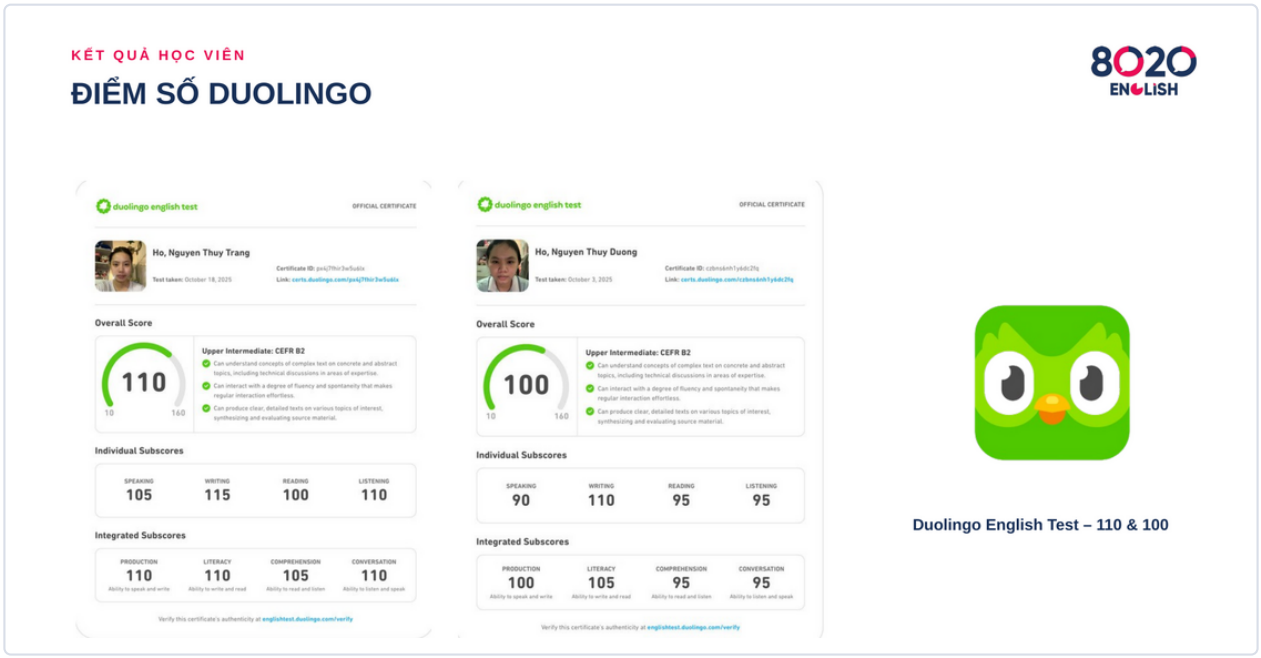
Điểm thi thật của học viên



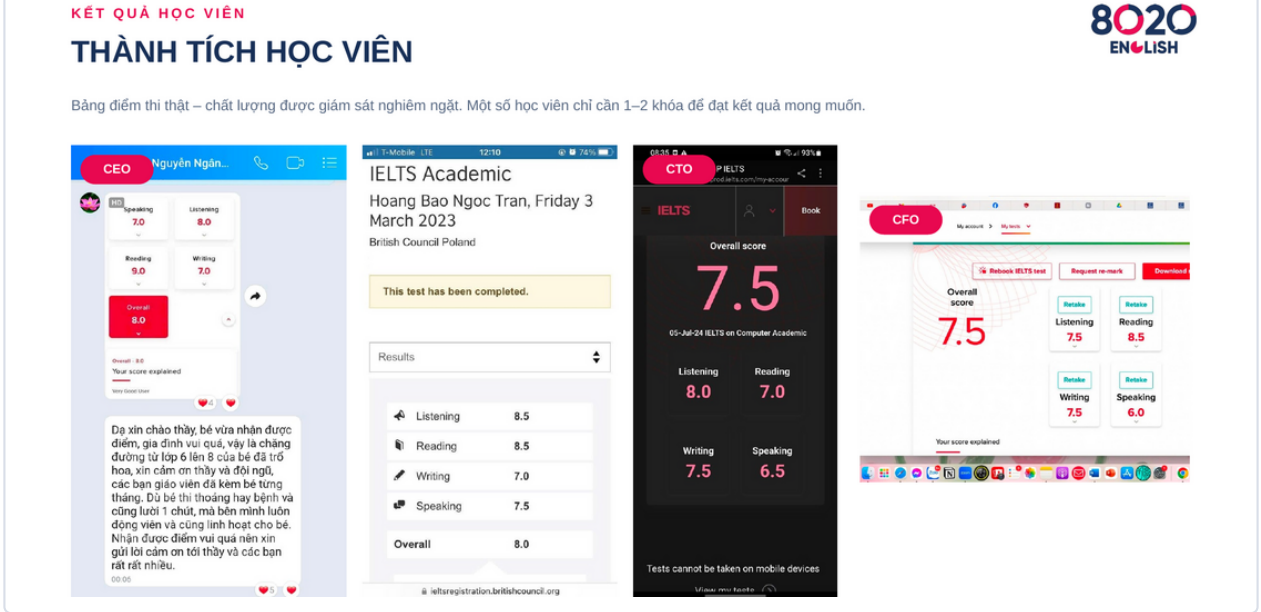
Bảng điểm AP thật — Calculus AB/BC 5/5 · Biology & Chemistry 4–5 (College Board)



Học viên 8020 — PTE Academic 90



Học viên 8020 — Duolingo English Test 110 & 100



Học viên 8020 — IELTS 8.0 / 7.5 / 8.5 (báo điểm thật)