

The Statistics of A/B Tests

The ABMeter Book

A short book on the mathematics needed to read the results of an online experiment: random variables, the central limit theorem, tests and intervals, randomization, and the delta method for ratios.

Every concept starts with the idea in plain words, then the definition or theorem, then worked examples with numbers small enough to check by hand.

Version 2026.10, built 2026-10-18

© ABMeter. This book is part of the ABMeter product and is free to read, copy and share.

Contents

	Preface	2
1	Random variables and estimators	3
	Random variables, mean and variance	3
	Covariance and independence	6
	Samples and estimators	8
2	Sampling distributions and the central limit theorem	13
	The sampling distribution and the law of large numbers	13
	Convergence in distribution and the central limit theorem	14
	How large is large	18
3	Hypothesis tests and confidence intervals	22
	Tests and p-values	22
	The two-sample z-test	25
	Confidence intervals	27
4	Randomized experiments	31
	Potential outcomes	31
	Assignment by hashing	34
5	Ratio metrics and the unit of analysis	37
	Mean of ratios and ratio of means	37
	The unit of analysis	38
	A ratio can move for the wrong reason	39
6	The delta method	40
	One variable	40
	Several means	42
	Computing the gradient	45
	Why the covariance matters	46
	Checking with the bootstrap	46
	Bias of a ratio	48
7	The relative lift	50
	The delta-method interval	50
	Fieller's interval	51
	Coverage	53
	Bias of the relative lift	53
8	Repeated looks, many comparisons and sample size	56
	Peeking	56
	Several comparisons	57
	Power and sample size	58
9	Statistics that are not means	61
	Formula sheet	63
	Glossary	64
	References	66
	Answers to the exercises	67

Preface

This book collects the statistics you need to read the result of an online experiment and to know why it is computed the way it is. It assumes you remember some calculus. It does not assume a statistics course.

The order is the usual one for the subject. Chapters 1 to 3 build the basic tools: random variables, the central limit theorem, tests and confidence intervals. Chapter 4 explains why a randomized experiment supports a causal conclusion, and how a hash function can stand in for a coin. Chapters 5 to 7 handle metrics that are not simple means, such as a ratio of two totals or a relative lift, with the delta method. Chapter 8 covers what goes wrong when you look at a test many times or run many tests at once, and how to choose a sample size. Chapter 9 is a short note on statistics that are not means at all.

How to read it

Each section opens with a short box, **The idea**, that says in plain words what the concept is for and gives a picture to hold in your head. Then come numbered blocks, as in most math books. A **Definition** fixes a term. A **Theorem** or **Proposition** states a fact, often followed by a short proof or proof sketch that shows why it holds. An **Example** works a toy case by hand with numbers small enough to check on paper. A **Remark** adds context or a caveat. Boxes marked **Naïve** and **Correct** put a common mistake and the right method side by side on the same data.

Every derivation shows each step on its own line, with the rule that justifies it on the right. When a symbol from an earlier section reappears, a short reminder of its meaning follows it in parentheses, so you can read any page without turning back. Numbers marked as coming from a simulation are produced by programs that run each time the book is built, with fixed random seeds, and the build checks that the text quotes them correctly. Every chapter ends with exercises; complete answers with the working are at the back.

Notation

The table is a reference only: every symbol is defined in the text where it first appears.

Symbol	Meaning
X, Y	random variables, for example one user's revenue and orders
$\mathbb{E}[X], \mu$	expectation (mean)
$\text{Var}(X), \sigma^2$	variance; σ is the standard deviation
$\text{Cov}(X, Y), \sigma_{XY}$	covariance
n	sample size: users in one group
$\bar{X}$	the sample mean of $X_1, \dots, X_n$
s^2, s_{XY}	sample variance and sample covariance (divided by $n - 1$)
SE	standard error: the standard deviation of an estimator
C, T	control and treatment group
Φ	the standard normal distribution function
z_q	the standard normal quantile, $\Phi(z_q) = q$; for example $z_{0.975} = 1.96$

In charts, blue is the correct method or the control group, orange is the naive method or the treatment group, and gray is a reference line.

Random variables and estimators

An experiment measures users, and users vary. This chapter sets up the language for that variation: random variables, their mean and variance, how two of them move together, and how to estimate all of this from a sample.

Random variables, mean and variance

The idea

Before a user arrives, you do not know what they will spend; you only know which amounts are possible and how often each happens. A random variable is that unknown number together with those odds. Its mean is the long-run average you would see over very many users. Its variance says how far a typical value sits from the mean: take each distance from the mean, square it so that values above and below do not cancel, and average.

Definition 1.1 (Random variable)

A **random variable** is a number whose value depends on chance. Its **distribution** says which values it takes and how often. We write $P(A)$ for the probability of an event A , for example $P(X = 1)$, and call $F(x) = P(X \leq x)$ the **distribution function** of X .

Example 1.2 (A checkout button)

A visitor reaches the checkout page and buys with probability 0.1. Let $X = 1$ if they buy and $X = 0$ if not. Then $P(X = 1) = 0.1$ and $P(X = 0) = 0.9$. The distribution function at 0.5 counts the values at or below 0.5, which is only the value 0: $F(0.5) = P(X \leq 0.5) = 0.9$.

Why squares, and why the square root

Try to measure spread with the plain deviations $X - \mu$. Their average is always 0: by linearity, $\mathbb{E}[X - \mu] = \mu - \mu = 0$, so the values above the mean cancel the values below. Something has to stop the cancelling. The two natural candidates are the absolute deviation $|X - \mu|$ and the squared deviation $(X - \mu)^2$, and for one variable either would do. The square won because of what happens with two. For independent X and Y the squared deviations add, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ ([Proposition 1.17](#)), the way squared legs add in Pythagoras' theorem; absolute deviations obey no such rule. That one property is what makes the spread of an average of n users computable ([Proposition 1.21](#)), and with it everything in Chapter 2. The price is the unit: the variance of a spend in dollars comes out in dollars squared, a number that cannot be compared with the mean or drawn on the same axis. The square root undoes the squaring, and the result, the standard deviation, is the spread back in the unit of the data.

Definition 1.3 (Expectation and variance)

The **expectation** (or **mean**) of a random variable X , written $\mathbb{E}[X]$, is its probability-weighted average:

$\mathbb{E}[X] = \sum_x x P(X = x)$ when X takes finitely many values, and the corresponding integral otherwise. We write $\mu = \mathbb{E}[X]$. The **variance** measures the spread around the mean:

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2]$$

We write $\sigma^2 = \text{Var}(X)$. The **standard deviation** (sd or SD in some tables) is its square root,

$$\sigma = \sqrt{\text{Var}(X)}$$

and has the same unit as X . When several variables are involved, a subscript names the variable: μ_X, σ_X^2, μ_Y .

Example 1.4 (Spend on a pricing page)

A visitor to a pricing page spends \$0 with probability 0.5, \$10 with probability 0.3 and \$30 with probability 0.2.

$$\begin{aligned} \mu &= 0 \times 0.5 + 10 \times 0.3 + 30 \times 0.2 && \text{(definition of the expectation)} \\ &= 0 + 3 + 6 = 9 && \text{(multiply and add)} \end{aligned}$$

The distances from the mean are -9 , 1 and 21 , and their squares are 81 , 1 and 441 .

$$\begin{aligned} \sigma^2 &= 81 \times 0.5 + 1 \times 0.3 + 441 \times 0.2 && \text{(definition of the variance)} \\ &= 40.5 + 0.3 + 88.2 = 129 && \text{(multiply and add)} \end{aligned}$$

The standard deviation is $\sqrt{129} = 11.4$ dollars: a typical visitor's spend is about \$11 away from the \$9 mean.

Proposition 1.5 (Linearity)

For constants a, b and random variables X, Y : $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$, and $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

Proof.

The first holds because an expectation is a weighted sum (or integral), and sums are linear. For the second, write $\mu = \mathbb{E}[X]$ (the mean of X):

$$\begin{aligned} \text{Var}(aX + b) &= \mathbb{E}[(aX + b - (a\mu + b))^2] && \text{(the mean of } aX + b \text{ is } a\mu + b, \text{ by the first part)} \\ &= \mathbb{E}[a^2(X - \mu)^2] && \text{(cancel } b, \text{ factor out } a) \\ &= a^2 \mathbb{E}[(X - \mu)^2] && \text{(linearity)} \\ &= a^2 \text{Var}(X) && \text{(definition of the variance)} \end{aligned}$$

Example 1.6 (Fees and currencies)

Take the pricing-page spend X of [Example 1.4](#), with mean $\mu = 9$ and variance $\sigma^2 = 129$. A report converts it to euros at 0.9 and adds a fixed 5-euro sign-up credit, $Y = 0.9X + 5$.

$$\begin{aligned} \mathbb{E}[Y] &= 0.9 \times 9 + 5 = 13.1 && \text{(linearity)} \\ \text{Var}(Y) &= 0.9^2 \times 129 = 104.49 && \text{(Proposition 1.5; the constant 5 drops out)} \end{aligned}$$

Adding the same amount to everyone moves the mean but not the spread.

Proposition 1.7 (A shortcut for the variance)

For a random variable X with mean $\mu = \mathbb{E}[X]$,

$$\text{Var}(X) = \mathbb{E}[X^2] - \mu^2$$

Proof.

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mu)^2] && \text{(definition of the variance)} \\ &= \mathbb{E}[X^2 - 2\mu X + \mu^2] && \text{(expand the square)} \\ &= \mathbb{E}[X^2] - 2\mu \mathbb{E}[X] + \mu^2 && \text{(linearity; } \mu \text{ is a constant)} \\ &= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 && (\mathbb{E}[X] = \mu) \\ &= \mathbb{E}[X^2] - \mu^2 && \text{(collect terms)}\end{aligned}$$

Example 1.8 (The shortcut on the pricing page)

For the spend of [Example 1.4](#), with $\mu = 9$ (the mean of one user's value):

$$\begin{aligned}\mathbb{E}[X^2] &= 0^2 \times 0.5 + 10^2 \times 0.3 + 30^2 \times 0.2 && \text{(definition of the expectation, applied to } X^2) \\ &= 0 + 30 + 180 = 210 && \text{(multiply and add)} \\ \text{Var}(X) &= 210 - 9^2 = 129 && \text{(the shortcut)}\end{aligned}$$

the same 129 as before, without computing any distances from the mean.

Example 1.9 (A 0/1 outcome)

Let $X = 1$ with probability p and $X = 0$ otherwise (a **Bernoulli** variable), like the checkout button above. Since X takes only the values 0 and 1, and $0^2 = 0$ and $1^2 = 1$, $X^2 = X$ for every outcome. For ten users with nine 0s and one 1, the squares are the same list, and both average 0.1. So

$$\begin{aligned}\mathbb{E}[X] &= 1 \times p + 0 \times (1 - p) = p && \text{(definition of the expectation)} \\ \mathbb{E}[X^2] &= \mathbb{E}[X] = p && (X^2 = X) \\ \text{Var}(X) &= \mathbb{E}[X^2] - p^2 && \text{(the shortcut, with } \mu = p) \\ &= p - p^2 = p(1 - p) && (\mathbb{E}[X^2] = p)\end{aligned}$$

A conversion rate of 10% has variance $0.1 \times 0.9 = 0.09$ per user.

The variance also bounds how often a variable strays far from its mean.

A bound that knows only the variance

Chebyshev wanted to say how often a variable strays far from its mean. The exact answer needs the whole distribution, which for a user's spend nobody has. His move was to ask what can be said from the variance alone, one number, and the answer is an inequality that is never wrong and rarely tight. Its job in this book is not to be sharp: it proves the law of large numbers in Chapter 2 with no assumption about the shape of the data.

Proposition 1.10 (Chebyshev's inequality)

For any random variable Y with mean m and variance v , and any distance $\varepsilon > 0$,

$$P(|Y - m| > \varepsilon) \leq \frac{v}{\varepsilon^2}$$

Proof.

Write $1\{A\}$ for the indicator of an event A : 1 when A happens and 0 otherwise.

$$\begin{aligned} v &= \mathbb{E}[(Y - m)^2] && \text{(definition of the variance)} \\ &\geq \mathbb{E}[(Y - m)^2 \cdot 1\{|Y - m| > \varepsilon\}] && \text{(drop the outcomes with } |Y - m| \leq \varepsilon; \text{ the rest is } \geq 0) \\ &\geq \mathbb{E}[\varepsilon^2 \cdot 1\{|Y - m| > \varepsilon\}] && ((Y - m)^2 > \varepsilon^2 \text{ on that event)} \\ &= \varepsilon^2 P(|Y - m| > \varepsilon) && \text{(the expectation of an indicator is a probability)} \end{aligned}$$

Dividing by ε^2 gives the inequality. ■

Example 1.11 (Chebyshev on the pricing page)

For the spend of [Example 1.4](#), the mean is $m = 9$ and the variance $v = 129$. How often is a visitor's spend more than \$20 from the mean?

$$P(|X - 9| > 20) \leq \frac{129}{20^2} = 0.3225$$

The exact answer: only \$30 is more than 20 away from 9, so the probability is 0.2. The bound holds and is loose, because it must hold for every distribution with that variance.

Covariance and independence

The idea

Covariance asks whether two quantities move together. For each user, take how far each quantity is from its own mean and multiply the two distances: when both are above or both are below, the product is positive. Averaged over users, a positive covariance means the two tend to rise and fall together. Independence is the stronger statement that knowing one tells you nothing at all about the other.

Why multiply the two deviations

A spread in one variable came from squaring a deviation. The same move for two variables is to multiply their deviations, $(X - \mu_X)(Y - \mu_Y)$. The product is positive when both values sit on the same side of their means and negative when they sit on opposite sides, so its average is positive when the two tend to rise together, negative when one rises as the other falls, and near 0 when the sides are unrelated. The variance is the special case $Y = X$. Multiplying also gives the result a unit, dollars times orders, and a size that depends on the scale of both variables, so a covariance of 0.10 means nothing on its own. Dividing by the two standard deviations removes both scales: that is the correlation, a pure number between -1 and 1 .

Definition 1.12 (Covariance and correlation)

For random variables X and Y with means μ_X and μ_Y , the **covariance** measures how they move together:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

We write $\sigma_{XY} = \text{Cov}(X, Y)$. The **correlation** $\sigma_{XY}/(\sigma_X\sigma_Y)$, with σ_X and σ_Y the standard deviations, rescales it to lie between -1 and 1 .

Example 1.13 (Opens and purchases in an email campaign)

For a user who received a campaign email, let $O = 1$ if they opened it and $B = 1$ if they bought. Half the users neither open nor buy, 30% open without buying, and 20% open and buy; nobody buys without opening. So $\mu_O = 0.5$ and $\mu_B = 0.2$ (the means of O and B).

O	B	probability	$(O - 0.5)(B - 0.2)$
0	0	0.5	$(-0.5)(-0.2) = 0.10$
1	0	0.3	$(0.5)(-0.2) = -0.10$
1	1	0.2	$(0.5)(0.8) = 0.40$

$$\begin{aligned}\sigma_{OB} &= 0.10 \times 0.5 - 0.10 \times 0.3 + 0.40 \times 0.2 && \text{(definition of the covariance)} \\ &= 0.05 - 0.03 + 0.08 = 0.10 && \text{(multiply and add)}\end{aligned}$$

With $\sigma_O = 0.5$ and $\sigma_B = 0.4$ (the square roots of the 0/1 variances 0.5×0.5 and 0.2×0.8 of [Example 1.9](#)), the correlation is $0.10 / (0.5 \times 0.4) = 0.5$. Openers buy more often, and the covariance is positive.

Example 1.14 (Zero correlation is not "unrelated")

Let X be $-1, 0$ or 1 with probability $1/3$ each, and $Y = X^2$, so Y is completely determined by X . Here $\mu_X = 0$ and $\mu_Y = 2/3$.

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}[X(Y - 2/3)] && \text{(definition of the covariance, } \mu_X = 0) \\ &= \frac{1}{3}((-1)(1 - \frac{2}{3}) + 0 + (1)(1 - \frac{2}{3})) && \text{(the three equally likely values)} \\ &= 0 && \text{(the two outer terms cancel)}\end{aligned}$$

Covariance and correlation measure only the straight-line part of a relationship, and a symmetric U shape has none, however strong it is. Nor does a positive correlation say which quantity drives the other: in the email example, people who already intend to buy may simply open more email. Chapter 4 shows how randomization separates cause from association.

Definition 1.15 (Independence)

X and Y are **independent** when knowing the value of one does not change the distribution of the other. For independent variables the expectation of a product is the product of expectations: $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$.

Example 1.16 (Two visitors, and an open and a purchase)

Two different visitors each buy with probability 0.1 , independently. The chance that both buy is $0.1 \times 0.1 = 0.01$, and indeed $\mathbb{E}[X_1 X_2] = 0.01 = \mathbb{E}[X_1] \mathbb{E}[X_2]$. In the email example, opening and buying are not independent: $\mathbb{E}[OB] = P(O = 1, B = 1) = 0.2$, while $\mathbb{E}[O] \mathbb{E}[B] = 0.5 \times 0.2 = 0.1$.

Proposition 1.17 (Variance of a sum)

For any random variables X and Y :

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$$

If X and Y are independent, $\text{Cov}(X, Y) = 0$ and the variances add.

Proof.

Write $\tilde{X} = X - \mu_X$ and $\tilde{Y} = Y - \mu_Y$, the deviations from μ_X and μ_Y (the means of X and Y).

$$\begin{aligned} \text{Var}(X + Y) &= \mathbb{E}[(\tilde{X} + \tilde{Y})^2] && \text{(the mean of } X + Y \text{ is } \mu_X + \mu_Y, \text{ by linearity)} \\ &= \mathbb{E}[\tilde{X}^2 + 2\tilde{X}\tilde{Y} + \tilde{Y}^2] && \text{(expand the square)} \\ &= \mathbb{E}[\tilde{X}^2] + 2\mathbb{E}[\tilde{X}\tilde{Y}] + \mathbb{E}[\tilde{Y}^2] && \text{(linearity)} \\ &= \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y) && \text{(definitions of variance and covariance)} \end{aligned}$$

For independent variables:

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[\tilde{X}\tilde{Y}] && \text{(definition of the covariance)} \\ &= \mathbb{E}[\tilde{X}] \mathbb{E}[\tilde{Y}] && \text{(independence)} \\ &= 0 && (\mathbb{E}[\tilde{X}] = \mathbb{E}[X] - \mu_X = 0) \end{aligned}$$

Example 1.18 (Actions per email user)

Count the actions of a user in [Example 1.13](#) as $O + B$: an open plus a purchase. With $\text{Var}(O) = 0.25$, $\text{Var}(B) = 0.16$ and $\sigma_{OB} = 0.10$ (the covariance of O and B):

$$\text{Var}(O + B) = 0.25 + 0.16 + 2 \times 0.10 = 0.61$$

Directly: $O + B$ is 0, 1 or 2 with probabilities 0.5, 0.3 and 0.2, so its mean is $0.3 + 0.4 = 0.7$, $\mathbb{E}[(O + B)^2]$ is $0.3 + 4 \times 0.2 = 1.1$, and the shortcut gives $1.1 - 0.49 = 0.61$. Ignoring the covariance would give 0.41, a third too small.

Samples and estimators

The idea

The distribution of a user's value is unknown; what you have is a sample, the users you actually observed. An estimator turns the sample into a guess, such as the average revenue of the users in the test. Run the experiment again with new users and the guess comes out a little different. The standard error measures how much it would wobble from run to run.

Definition 1.19 (Sample, estimator, bias, consistency)

A **sample** $X_1, \dots, X_n$ is n independent draws from the same distribution; such variables are called **independent and identically distributed**, abbreviated **i.i.d.** An **estimator** is a function of the sample that is meant to approximate an unknown quantity θ ; a hat marks the estimator, $\hat{\theta}$. Its **bias** is $\mathbb{E}[\hat{\theta}] - \theta$; it is **unbiased** if the bias is 0, and **consistent** if $\hat{\theta}$ approaches θ as n grows.

Example 1.20 (Two signups, two estimators)

A signup form converts with probability $p = 0.2$, and two visitors arrive. The observed rate $\hat{p}$ (signups divided by 2) is 0, 0.5 or 1, with probabilities $0.8 \times 0.8 = 0.64$, $2 \times 0.2 \times 0.8 = 0.32$ and $0.2 \times 0.2 = 0.04$.

$$\mathbb{E}[\hat{p}] = 0 \times 0.64 + 0.5 \times 0.32 + 1 \times 0.04 = 0.2$$

so $\hat{p}$ is unbiased. A “smoothed” rate, $(\text{signups} + 1)/(2 + 2)$, takes the values 0.25, 0.5 and 0.75 with the same probabilities, so its expectation is $0.16 + 0.16 + 0.03 = 0.35$: a bias of 0.15.

Why the error shrinks like $\sqrt{n}$ and not like n

Average two users whose spends have the same variance σ^2 . If both happen to err in the same direction, the average errs as much as one user does; if they err in opposite directions, the errors cancel. Independence makes neither case the rule, so a user’s error neither survives whole nor vanishes in the average: by [Proposition 1.17](#) the variance of the sum is $2\sigma^2$, and dividing the sum by 2 divides the variance by 4 ([Proposition 1.5](#)), which leaves $\sigma^2/2$. Each further independent user divides the variance of the average by the next integer. The standard deviation, the thing in the data’s unit, falls only as the square root of that: four times the users halve the error, and a hundred times the users divide it by ten. That is the arithmetic behind every sample-size question in Chapter 8, and it is why more data is expensive.

Proposition 1.21 (The sample mean)

Let $X_1, \dots, X_n$ be i.i.d. with mean $\mu = \mathbb{E}[X_i]$ and variance $\sigma^2 = \text{Var}(X_i)$. The **sample mean** $\bar{X} = (X_1 + \dots + X_n)/n$ (a bar marks a sample mean) satisfies

$$\mathbb{E}[\bar{X}] = \mu$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Proof.

$$\begin{aligned} \mathbb{E}[\bar{X}] &= (\mathbb{E}[X_1] + \dots + \mathbb{E}[X_n])/n && \text{(linearity)} \\ &= n\mu/n = \mu && (\mathbb{E}[X_i] = \mu) \end{aligned}$$

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}(X_1 + \dots + X_n)/n^2 && \text{(Proposition 1.5 with } a = 1/n) \\ &= (\text{Var}(X_1) + \dots + \text{Var}(X_n))/n^2 && \text{(independent, so variances add: Proposition 1.17)} \\ &= n\sigma^2/n^2 = \sigma^2/n && (\text{Var}(X_i) = \sigma^2) \end{aligned}$$

Example 1.22 (The average of 100 pricing-page visitors)

For the spend of [Example 1.4](#), $\mu = 9$ and $\sigma^2 = 129$ (the mean and the variance of one user’s value). The average of 100 visitors has mean 9 and variance $129 / 100 = 1.29$, so its standard deviation is $\sqrt{1.29} = 1.14$ dollars: ten times smaller than one visitor’s 11.4.

Definition 1.23 (Standard error)

The **standard error** of an estimator, written SE, is its standard deviation. For the sample mean of n i.i.d. observations with standard deviation σ it is $\sigma/\sqrt{n}$, by [Proposition 1.21](#), and since σ is unknown it is estimated by

$$SE = \frac{s}{\sqrt{n}}$$

where s^2 is the **sample variance**:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

and s is the **sample standard deviation**. For pairs (X_i, Y_i) the **sample covariance** s_{XY} is defined in the same way, with $(X_i - \bar{X})(Y_i - \bar{Y})$ in place of the square, and s_X^2, s_Y^2 name the sample variances of X and Y .

Example 1.24 (Five users, by hand)

Revenue per user is 0, 0, 20, 0, 40. The mean is $60 / 5 = 12$. The squared deviations from 12 are 144, 144, 64, 144 and 784, which add up to 1280, so

$$s^2 = 1280/4 = 320 \quad (\text{definition of the sample variance})$$

$$SE = \sqrt{320/5} = 8 \quad (SE = s/\sqrt{n})$$

For the same users as a 0/1 outcome (bought or not), the observed share of buyers is $\hat{p} = 2/5 = 0.4$, and the shortcut $s^2 = n \hat{p}(1 - \hat{p})/(n - 1)$ gives $5 \times 0.4 \times 0.6 / 4 = 0.3$. The shortcut holds because, for 0/1 data,

$$\begin{aligned} \sum_i (X_i - \hat{p})^2 &= \sum_i X_i^2 - 2\hat{p} \sum_i X_i + n\hat{p}^2 && (\text{expand the square}) \\ &= n\hat{p} - 2n\hat{p}^2 + n\hat{p}^2 && (X_i^2 = X_i \text{ and } \sum_i X_i = n\hat{p}) \\ &= n\hat{p}(1 - \hat{p}) && (\text{collect terms}) \end{aligned}$$

and dividing by $n - 1$ gives s^2 .

Example 1.25 (Revenue and orders)

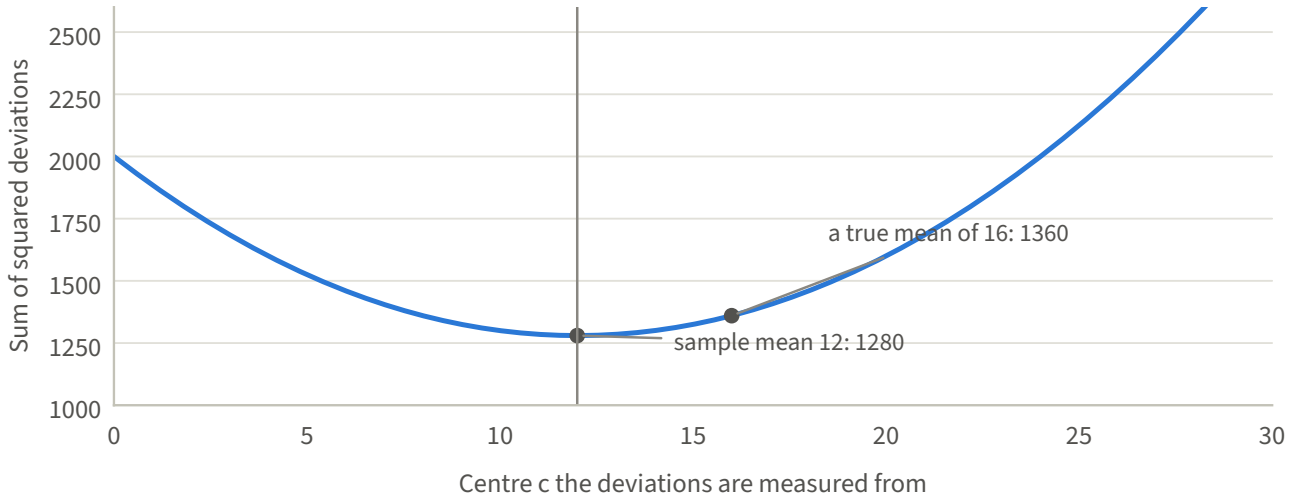
The same five users placed $Y = (0, 0, 1, 0, 3)$ orders, so revenue and orders come in pairs (X_i, Y_i) . Their sample variances are $s_X^2 = 320$ and $s_Y^2 = 1.7$ (the orders have mean 0.8 and squared deviations 0.64, 0.64, 0.04, 0.64, 4.84, summing to 6.8), and the sample covariance is $s_{XY} = 23$ (products of deviations $(-12)(-0.8) \times 3 + 8 \times 0.2 + 28 \times 2.2 = 92$, divided by 4). The correlation is $23 / 23.3 = 0.986$, where 23.3 is the square root of 320×1.7 : users with more orders spend more. For the sample variance of $X + Y$:

$$\begin{aligned} s_{X+Y}^2 &= s_X^2 + s_Y^2 + 2s_{XY} && (\text{Proposition 1.17, sample version}) \\ &= 320 + 1.7 + 46 = 367.7 && (\text{substitute}) \end{aligned}$$

which matches the sample variance computed directly from the five sums. Without the covariance it would be 321.7.

Why $n - 1$: the sample mean sits too close to the data

The variance is the average squared distance from μ , the true mean. The sample does not contain μ ; it contains $\bar{X}$, computed from the same values, and $\bar{X}$ sits closer to the sample than μ does: of all the centres c you could measure deviations from, $\bar{X}$ is the one that makes $\sum_i (X_i - c)^2$ smallest, as the figure below shows. So the deviations from $\bar{X}$ are systematically a little too small. In the five-user sample above, if the true mean were 16 rather than the sample's 12, the squared deviations from 16 would add to 1360, not 1280. The shortfall is not noise that averages out; it is always in the same direction, and on average it is exactly one observation's worth of spread, σ^2 . Dividing by $n - 1$ instead of n puts it back. With $n = 1$ the formula gives $0/0$ and refuses to answer, which is right: one value says nothing about spread.



The sum of squared deviations of the five values 0, 0, 20, 0, 40 as a function of the centre they are measured from. It is lowest, 1280, at the sample mean 12. Whatever the true mean is, it is somewhere else on the curve, so the true squared deviations add up to more, here 1360 at 16.

Proposition 1.26 (Why $n - 1$)

Let $X_1, \dots, X_n$ be i.i.d. with mean μ and variance σ^2 . The sample variance s^2 is unbiased: $\mathbb{E}[s^2] = \sigma^2$. Dividing by n instead gives $(n - 1)\sigma^2/n$ on average.

Proof.

Write $X_i - \bar{X} = (X_i - \mu) - (\bar{X} - \mu)$, with $\bar{X}$ the sample mean and μ the mean of one user's value. Squaring and summing over i :

$$\begin{aligned} \sum_i (X_i - \bar{X})^2 &= \sum_i (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_i (X_i - \mu) + n(\bar{X} - \mu)^2 && \text{(expand the square)} \\ &= \sum_i (X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2 && (\sum_i (X_i - \mu) = n(\bar{X} - \mu)) \\ &= \sum_i (X_i - \mu)^2 - n(\bar{X} - \mu)^2 && \text{(collect terms)} \end{aligned}$$

Take expectations, with σ^2 the variance of one user's value:

$$\begin{aligned} \mathbb{E}\left[\sum_i (X_i - \bar{X})^2\right] &= \sum_i \mathbb{E}[(X_i - \mu)^2] - n \mathbb{E}[(\bar{X} - \mu)^2] && \text{(linearity)} \\ &= n\sigma^2 - n \text{Var}(\bar{X}) && \text{(definition of the variance; } \mathbb{E}[\bar{X}] = \mu) \\ &= n\sigma^2 - n \cdot \sigma^2/n && \text{(Proposition 1.21)} \\ &= (n - 1)\sigma^2 && \text{(collect terms)} \end{aligned}$$

Dividing by $n - 1$ gives $E[s^2] = \sigma^2$. The deviations are measured from $\bar{X}$, which is fitted to the same data, and that costs one observation's worth of spread.

Example 1.27 (Two users, every possible sample)

A user's value is 0 or 2 with probability $\frac{1}{2}$ each, so $\mu = 1$ and $\sigma^2 = 1$ (the mean and the variance of one user's value). With $n = 2$ there are four equally likely samples:

sample	$\bar{X}$	$\sum_i (X_i - \bar{X})^2$	divided by $n - 1 = 1$	divided by $n = 2$
0, 0	0	0	0	0
0, 2	1	2	2	1
2, 0	1	2	2	1
2, 2	2	0	0	0

Averaged over the four samples, dividing by $n - 1$ gives $(0 + 2 + 2 + 0)/4 = 1$, the true variance; dividing by n gives 0.5, half of it. A larger check: a population of ten users spends 0, 0, 0, 0, 0, 0, 0, 20, 40 and 100, so $\mu = 16$ and $\sigma^2 = 9440/10 = 944$ (the squared deviations are 256 seven times, 16, 576 and 7056). In 200,000 simulated samples of five users drawn with replacement, dividing by n averaged 758.3, close to the expected $(4/5) \times 944 = 755.2$, and dividing by $n - 1$ averaged 947.9, close to 944. At 1,000 users the bias of dividing by n is only 1/1000, 0.1%.

In ABMeter, the standard error shown next to a metric that is a single per-user mean, such as a conversion rate or revenue per user, is this $s/\sqrt{n}$.

Exercises

1. A 0/1 outcome takes the values 1, 0, 0, 1. Compute $\bar{X}$, s^2 and SE (the sample mean, the sample variance and the standard error), and check s^2 with the Bernoulli shortcut.
2. For the revenue and orders of the example above, $s_X^2 = 320$, $s_Y^2 = 1.7$ and $s_{XY} = 23$. Compute the sample variance of $X - Y$ without recomputing from the data.
3. A sample of 400 users has a standard error of \$2 for mean revenue. How many users from the same population give a standard error of \$1?

Sampling distributions and the central limit theorem

A sample mean is itself a random variable: run the experiment again with new users and you get a different mean. This chapter describes how that mean is distributed. Two theorems do the work. The law of large numbers says the mean settles at the true value. The central limit theorem says how it fluctuates on the way, and that the fluctuations are approximately normal whatever the data look like.

The sampling distribution and the law of large numbers

The idea

Imagine re-running the same experiment many times, each time with fresh users, and writing down the average each time. Those averages form a distribution of their own, the sampling distribution, and its spread is the standard error. The law of large numbers says that with more users per run, the averages crowd ever closer to the true mean.

Definition 2.1 (Sampling distribution)

The **sampling distribution** of an estimator is its distribution over repeated samples of the same size from the same population. Its standard deviation is the standard error.

Example 2.2 (The sampling distribution of a signup rate)

A signup form converts with probability $p = 0.2$, and each run of the experiment sees two visitors. As in Chapter 1, the observed rate $\hat{p}$ is 0, 0.5 or 1 with probabilities 0.64, 0.32 and 0.04: that list is its sampling distribution, with mean 0.2.

$$\begin{aligned}\mathbb{E}[\hat{p}^2] &= 0.25 \times 0.32 + 1 \times 0.04 = 0.12 && \text{(definition of the expectation)} \\ \text{Var}(\hat{p}) &= 0.12 - 0.2^2 = 0.08 && \text{(Proposition 1.7)} \\ \text{SE} &= \sqrt{0.08} = 0.283 && \text{(standard deviation of the sampling distribution)}\end{aligned}$$

which matches [Proposition 1.21](#), with σ the standard deviation of one user's value:

$$\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{0.2 \times 0.8}{2}} = 0.283$$

Theorem 2.3 (Law of large numbers)

Let $X_1, X_2, \dots$ be i.i.d. with mean $\mu = \mathbb{E}[X_i]$ and finite variance $\sigma^2 = \text{Var}(X_i)$, and let $\bar{X}_n$ be the mean of the first n of them. Then for every distance $\varepsilon > 0$, the probability that $\bar{X}_n$ is more than ε away from μ goes to 0 as n grows:

$$P(|\bar{X}_n - \mu| > \varepsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

Proof.

$\bar{X}_n$ has mean μ (the mean of one user's value) and variance σ^2/n , with σ^2 the variance of one user's value (Proposition 1.21). By Chebyshev's inequality (Proposition 1.10) applied to $\bar{X}_n$, for every distance $\varepsilon > 0$:

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n \varepsilon^2}$$

The right side does not depend on the data, and

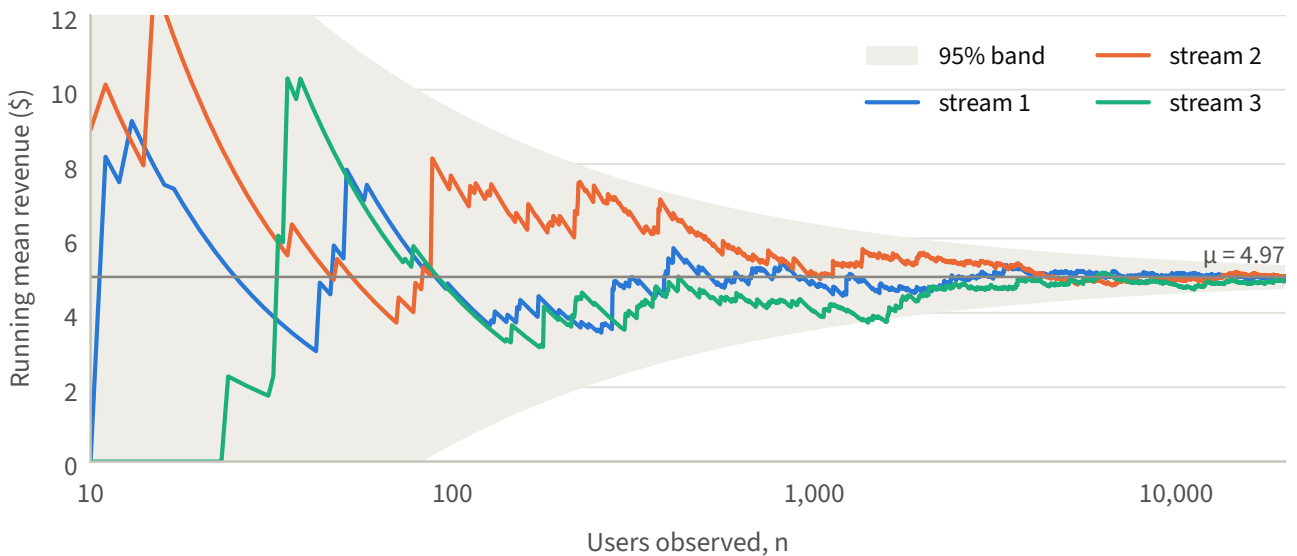
$$\frac{\sigma^2}{n \varepsilon^2} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

so the probability on the left goes to 0 as well. ■

Example 2.4 (How loose the bound is)

Take $\sigma = 10$ (the standard deviation of one user's value) and $\varepsilon = 1$ (the allowed distance from the mean). The bound $\sigma^2/(n \varepsilon^2) = 100/n$ is 1 at $n = 100$, which says nothing, since every probability is at most 1; 0.01 at $n = 10,000$; and 0.0001 at $n = 1,000,000$. The bound holds for every distribution, which makes it loose: the central limit theorem of the next section gives far smaller probabilities for the same n , as Exercise 3 shows.

The bound is crude, but it shows the mechanism: the variance of the mean falls like $1/n$. The chart follows the running mean of three independent streams of skewed revenue data, where 90% of users spend nothing.



Running means of three independent streams of revenue per user. Each converges to the true mean μ ; the gray band is where the central limit theorem of the next section expects 95% of means of n users.

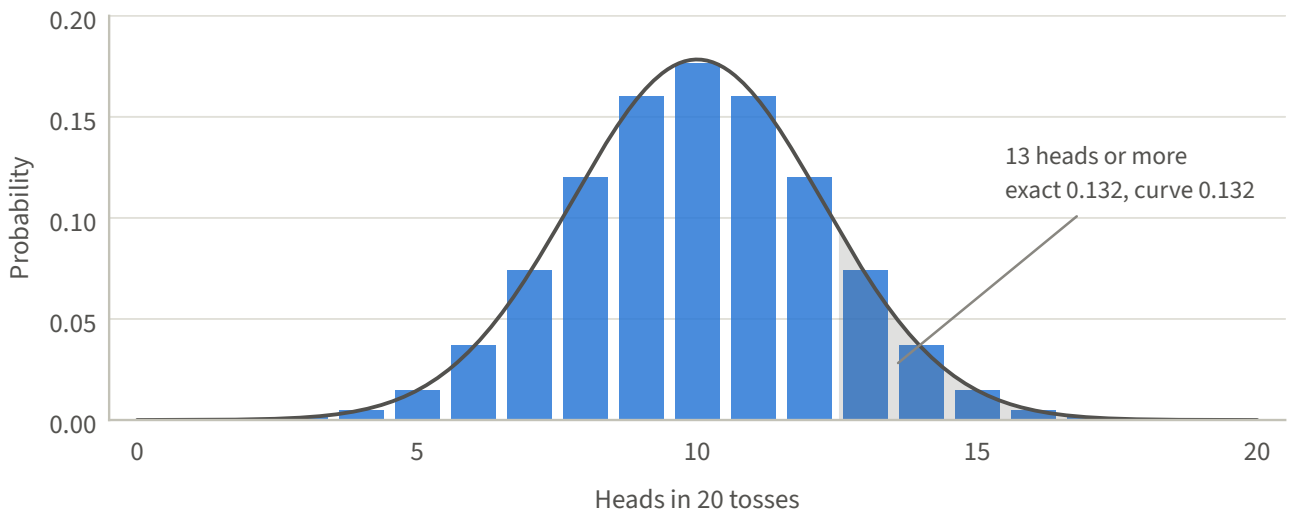
Convergence in distribution and the central limit theorem

The idea

The law of large numbers says where $\bar{X}$ ends up, not what its error looks like on the way. Blow the error up by $\sqrt{n}$, so that it stops shrinking, and draw its histogram over many re-runs. As n grows, that histogram settles into one fixed shape, and the shape is always the same bell, the normal distribution, whatever a single user's values look like. "Converges in distribution" is the precise name for "the histogram settles".

Where the bell came from

In 1733 Abraham de Moivre needed the probability that a fair coin tossed n times lands heads between a and b times, a question gamblers paid for. The exact answer is a sum of binomial terms, and for n in the hundreds the factorials in each term were beyond hand calculation. He found that for large n the whole sum is close to the area under one smooth curve, $e^{-x^2/2}$ after centring and scaling, and a single curve can be tabulated once and for all. Laplace and Gauss then met the same curve wherever many small independent errors are added, first in astronomical measurements, and a century later in sums of anything with a finite variance. The bell is not a fact about users or coins. It is what adding many independent contributions produces, whatever each contribution looks like. In the formula below, the constant in front only makes the total area 1, and the $2\sigma^2$ in the exponent places the two points where the curve changes from convex to concave exactly one standard deviation from the mean.



De Moivre's problem. Bars: the exact probabilities of each number of heads in 20 fair tosses. Curve: the normal density with the same mean, 10, and standard deviation, $\sqrt{5} = 2.24$. Its area between 12.5 and 20 is 0.132, against the exact 0.132 for 13 heads or more.

Definition 2.5 (Normal distribution)

The **normal distribution** $N(\mu, \sigma^2)$ with mean μ and variance σ^2 has the bell-shaped density

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

We write $X \sim N(\mu, \sigma^2)$ for “ X has this distribution”. $N(0, 1)$ is the **standard normal** distribution, and its distribution function is written $\Phi(x) = P(Z \leq x)$ for $Z \sim N(0, 1)$. If $X \sim N(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim N(0, 1)$. The **quantile** z_q is the value with $\Phi(z_q) = q$; for example $z_{0.975} = 1.96$, so a standard normal variable lies within ± 1.96 with probability 95%.

Example 2.6 (A high average on the pricing page)

Suppose the average spend $\bar{X}$ of 100 pricing-page visitors is normal with mean 9 and variance 1.29, as for the spend of Chapter 1 (the central limit theorem below says this is approximately true). How likely is an average above \$11? With Z a standard normal variable and Φ the standard normal distribution function, $\Phi(1.76) = 0.9608$:

$$\begin{aligned}
 P(\bar{X} > 11) &= P\left(\frac{\bar{X} - 9}{\sqrt{1.29}} > \frac{11 - 9}{\sqrt{1.29}}\right) && \text{(subtract the mean and divide by the sd on both sides)} \\
 &= P(Z > 1.76) && \text{(Definition 2.5: the standardized variable is } N(0, 1)) \\
 &= 1 - \Phi(1.76) && \text{(complement)} \\
 &= 1 - 0.9608 = 0.039 && \text{(table value)}
 \end{aligned}$$

Why compare probabilities and not values

The obvious meaning of “ Z_n converges to Z ” is that the numbers get close. For the rescaled error that meaning is useless: every re-run of the experiment produces a fresh error, unrelated to the last one, so the values never settle. What settles is the histogram, and a histogram is the same information as the list of probabilities $P(Z_n \leq x)$ for every x . So the definition compares those probabilities and asks nothing of the values themselves.

Definition 2.7 (Convergence in distribution)

A sequence of random variables $Z_1, Z_2, \dots$ **converges in distribution** to a random variable Z if

$$P(Z_n \leq x) \rightarrow P(Z \leq x) \quad \text{as } n \rightarrow \infty$$

for every x at which the distribution function of Z is continuous. This is written

$$Z_n \xrightarrow{d} Z$$

and read aloud as “ Z_n converges in distribution to Z ”. When the limit is written as a distribution, as in the central limit theorem below, it means a random variable with that distribution.

Example 2.8 (A finer and finer grid)

Let Z_n be one of the values $1/n, 2/n, \dots, n/n$, each with probability $1/n$, and let U be uniform on $[0, 1]$: $P(U \leq x) = x$. At $x = 0.3$, $P(Z_n \leq 0.3)$ counts the grid points at or below 0.3:

n	grid points ≤ 0.3	$P(Z_n \leq 0.3)$
4	1/4	0.25
10	1/10, 2/10, 3/10	0.30
100	1/100 to 30/100	0.30

In general $P(Z_n \leq x)$ is $\lfloor nx \rfloor / n$ (with $\lfloor \cdot \rfloor$ rounding down), which is within $1/n$ of x , so $Z_n \xrightarrow{d} U$. No single value of Z_n approaches any value of U ; only the histograms become alike.

The definition compares probabilities, not values. A normal distribution function is continuous everywhere, so for a normal limit the condition holds at every x .

Why multiply by $\sqrt{n}$

Consider what a theorem could say about $\bar{X}_n - \mu$. Its variance is σ^2/n , so the error itself converges to 0, and a limit of 0 has no shape to describe. Multiply by n instead and the variance becomes $n\sigma^2$, which blows up. Between the two, exactly one magnification holds the variance fixed: multiplying by $\sqrt{n}$ gives variance $n \cdot \sigma^2/n = \sigma^2$ ([Proposition 1.5](#)). At that magnification the error neither shrinks nor grows, so its histogram can settle into one shape, and the theorem can name it.

Theorem 2.9 (Central limit theorem)

Let $X_1, \dots, X_n$ be i.i.d. with mean $\mu = \mathbb{E}[X_i]$ and finite variance $\sigma^2 = \text{Var}(X_i) > 0$, and let $\bar{X}_n$ be their mean. Then

$$\sqrt{n} \cdot (\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Equivalently, dividing by σ ([Definition 2.5](#)), for every x ,

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) \rightarrow \Phi(x)$$

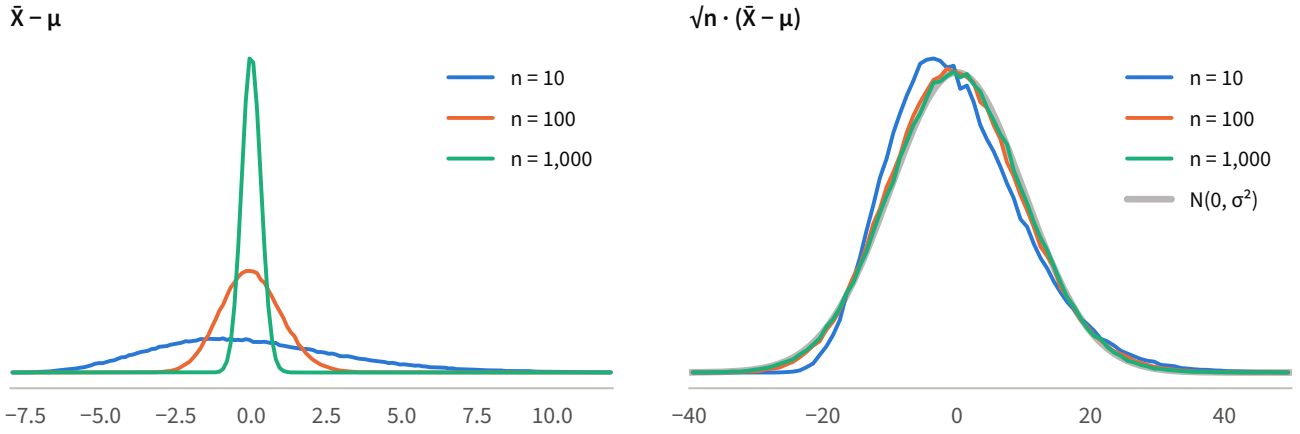
Example 2.10 (One bell, two sizes)

Revenue per user has standard deviation $\sigma = 10$. With $n = 100$ users, $\sigma/\sqrt{n} = 10/10 = 1$, so $\bar{X} - \mu$ is approximately $N(0, 1)$ and $\bar{X}$ lies within about ± 2 (precisely ± 1.96) of μ (the mean of one user's value) in 95% of samples. With $n = 10,000$, $\sigma/\sqrt{n} = 10/100 = 0.1$ and the band is $\pm 2 \times 0.1 = \pm 0.2$. In both cases the rescaled error $\sqrt{n} \cdot (\bar{X} - \mu)$ has approximately the same distribution, $N(0, \sigma^2) = N(0, 100)$.

A simulation checks this for skewed data: exponential revenue with mean and standard deviation 10, 20,000 samples per size.

n	middle 95% of $\bar{X} - \mu$	$1.96 \sigma/\sqrt{n}$	sd of $\sqrt{n} \cdot (\bar{X} - \mu)$
10	-5.25 to 7.06	6.20	10.01
100	-1.86 to 2.06	1.96	10.05
1,000	-0.61 to 0.64	0.62	10.06
10,000	-0.19 to 0.19	0.20	9.93

The middle 95% shrinks tenfold for every hundredfold increase in n (at $n = 10$ it is still lopsided, because the data are skewed), while the standard deviation of the rescaled error stays at about 10.



Simulation, 20,000 samples per size. Left: the error of the sample mean for $n = 10, 100$ and $1,000$ collapses towards 0. Right: the same errors multiplied by $\sqrt{n}$ settle into one bell, the normal distribution with standard deviation $\sigma = 10$ (gray curve).

Remark 2.11 (Reading the theorem in words)

The error $\bar{X}_n - \mu$, with μ the mean of one user's value, shrinks to 0 like $1/\sqrt{n}$: its standard deviation is $\sigma/\sqrt{n}$, with σ the standard deviation of one user's value ([Proposition 1.21](#)). Multiplying by $\sqrt{n}$ cancels the shrinking:

$$\begin{aligned} \text{Var}(\sqrt{n} \cdot (\bar{X}_n - \mu)) &= n \text{Var}(\bar{X}_n) && \text{(Proposition 1.5 with } a = \sqrt{n}\text{)} \\ &= n \cdot \sigma^2/n && \text{(Proposition 1.21)} \\ &= \sigma^2 && \text{(cancel } n\text{)} \end{aligned}$$

So the rescaled error neither shrinks nor grows; the theorem says it settles into a fixed bell, the normal distribution with standard deviation σ . Dividing back by $\sqrt{n}$, which divides the variance by n ([Proposition 1.5](#)), gives the practical reading: for large n , $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$, and that is where the standard error $\sigma/\sqrt{n}$ comes from.

The theorem holds with the sample standard deviation s in place of σ (the standard deviation of one user's value), because s converges to σ (Slutsky's theorem; Wasserman 2004, Chapter 5). That version is the one used in practice: $(\bar{X} - \mu)/\text{SE}$, with $\text{SE} = s/\sqrt{n}$ the standard error, is approximately standard normal for large n .

How large is large

The idea

The central limit theorem promises a bell eventually, but not by when. The answer depends on how lopsided a single user's values are. A symmetric metric looks normal after a few dozen users; revenue where most users spend nothing and a few spend a lot needs thousands, because the rare big spenders keep tilting the average.

Theorem 2.12 (Berry–Esseen bound)

Let $X_1, \dots, X_n$ be i.i.d. with mean μ , standard deviation σ and $\rho = \mathbb{E}[|X_i - \mu|^3] < \infty$. The error of the normal approximation, measured as the largest gap over all x ($\sup_x$) between the two distribution functions, is at most

$$\sup_x \left| P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) - \Phi(x) \right| \leq \frac{C \rho}{\sigma^3 \sqrt{n}}$$

for a universal constant C below 0.5.

Example 2.13 (Berry-Esseen for a 10% conversion rate)

For a 0/1 outcome with $p = 0.1$, the mean is 0.1 and $\sigma = 0.3$, the square root of 0.1×0.9 (the standard deviation of one user's value).

$$\begin{aligned}\rho &= 0.1 \times 0.9^3 + 0.9 \times 0.1^3 && (|X - 0.1| \text{ is } 0.9 \text{ or } 0.1) \\ &= 0.0729 + 0.0009 = 0.0738 && (\text{multiply and add}) \\ \frac{\rho}{\sigma^3} &= 0.0738 / 0.027 = 2.73 && (\sigma^3 = 0.3^3)\end{aligned}$$

With $C = 0.5$ the bound is $0.5 \times 2.73 / \sqrt{n}$: 0.137 at $n = 100$ and 0.0137 at $n = 10,000$. At 100 users a probability read off the normal curve could, in the worst case, be off by 14 points; at 10,000 users by at most 1.4.

Why the cube

The variance squares deviations, which erases their sign: a long tail to the right and a long tail to the left give the same variance. To tell them apart the sign must survive, so the next odd power is used, $(X - \mu)^3$. Large deviations still dominate, as with the square, but the ones above the mean count as positive and the ones below as negative, and a symmetric distribution comes out at 0. Dividing by σ^3 removes the unit, so that revenue in dollars and the same revenue in cents have the same skewness, and the skewness of different metrics can be compared.

Definition 2.14 (Skewness)

For a random variable X with mean μ and standard deviation σ , the **third central moment** is $\mathbb{E}[(X - \mu)^3]$, and the **skewness** is $\gamma = \mathbb{E}[(X - \mu)^3] / \sigma^3$. It is 0 for a symmetric distribution and positive when the distribution has a long right tail. It has no unit, so rescaling X does not change it.

Example 2.15 (Skewness of a 0/1 outcome)

A Bernoulli variable X with $P(X = 1) = p$ has mean p and variance $\sigma^2 = p(1 - p)$.

$$\begin{aligned}\mathbb{E}[(X - p)^3] &= p(1 - p)^3 + (1 - p)(0 - p)^3 && (X = 1 \text{ with probability } p, \text{ else } 0) \\ &= p(1 - p)[(1 - p)^2 - p^2] && (\text{factor out } p(1 - p)) \\ &= p(1 - p)(1 - 2p) && ((1 - p)^2 - p^2 = 1 - 2p) \\ &= \sigma^2(1 - 2p) && (\sigma^2 = p(1 - p)) \\ \gamma &= \sigma^2(1 - 2p) / \sigma^3 = (1 - 2p) / \sigma && (\text{definition of the skewness})\end{aligned}$$

The skewness is 0 at $p = 0.5$, $0.8 / 0.3 = 2.67$ at $p = 0.1$ and about 9.9 at $p = 0.01$. Rare events are as lopsided as skewed revenue.

Proposition 2.16 (Skewness of the mean)

Let $X_1, \dots, X_n$ be i.i.d. with mean μ , standard deviation σ and skewness γ , and let $\bar{X}_n$ be their mean. The skewness of $\bar{X}_n$ is $\gamma / \sqrt{n}$.

Proof.

Let $S = X_1 + \dots + X_n$ and $\tilde{X}_i = X_i - \mu$, so $S - n\mu = \tilde{X}_1 + \dots + \tilde{X}_n$. In the expansion of the cube, every cross term contains some $\tilde{X}_j$ to the first power times factors independent of it:

$$\begin{aligned}\mathbb{E}[(S - n\mu)^3] &= \sum_i \mathbb{E}[\tilde{X}_i^3] && \text{(cross terms have expectation } \mathbb{E}[\tilde{X}_j] \cdot (\dots) = 0 \text{ by independence)} \\ &= n\gamma\sigma^3 && \text{(definition of the skewness: } \mathbb{E}[\tilde{X}_i^3] = \gamma\sigma^3\text{)}\end{aligned}$$

With $\text{Var}(S) = n\sigma^2$ ([Proposition 1.17](#)), the skewness of S is

$$\begin{aligned}\frac{\mathbb{E}[(S - n\mu)^3]}{\text{Var}(S)^{3/2}} &= \frac{n\gamma\sigma^3}{(n\sigma^2)^{3/2}} && \text{(the two results above)} \\ &= \frac{n\gamma\sigma^3}{n^{3/2}\sigma^3} && ((n\sigma^2)^{3/2} = n^{3/2}\sigma^3) \\ &= \frac{\gamma}{\sqrt{n}} && \text{(cancel } n\sigma^3\text{)}\end{aligned}$$

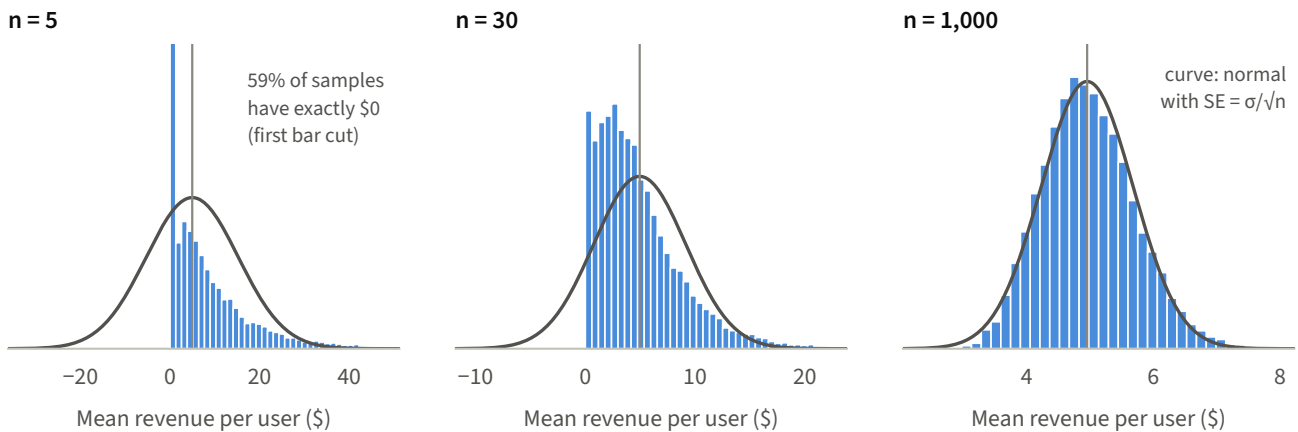
$\bar{X}_n = S/n$ has the same skewness: dividing by n divides the third central moment by n^3 and the standard deviation by n ([Proposition 1.5](#)), so its cube also by n^3 . ■

Example 2.17 (Skewed revenue)

Take revenue per user in which 90% of users spend nothing and a buyer spends a lognormal amount (a variable whose logarithm is normal, here with log-mean 3.5 and log-standard deviation 0.9, so a median purchase of $e^{3.5} = \$33$). Its exact mean is $0.1 \times e^{3.5+0.9^2/2} = 4.97$; estimated from five million simulated users, the mean is 4.95, the standard deviation $\sigma = 22.98$ and the skewness $\gamma = 10.6$. By the proposition, the mean of 30 users still has skewness $10.6 / \sqrt{30} \approx 10.6 / 5.48 \approx 1.94$, and the mean of 1,000 users has $10.6 / \sqrt{1000} \approx 10.6 / 31.6 \approx 0.34$. A simulation of 20,000 samples of each size measures the spread, the skewness, and how often the mean lands more than 1.96 standard errors above or below μ ; a normal distribution would give 2.5% on each side.

n	sd of the mean	$\sigma/\sqrt{n}$	skewness	$\gamma/\sqrt{n}$	above +1.96 SE	below -1.96 SE
5	9.98	10.28	4.33	4.76	4.3%	0.0%
30	4.18	4.20	1.96	1.94	4.5%	0.0%
1,000	0.735	0.727	0.36	0.34	3.6%	1.4%

The spread matches $\sigma/\sqrt{n}$ at every size, because [Proposition 1.21](#) is exact. The shape converges slowly: at $n = 30$ the mean almost never falls 1.96 standard errors below μ and does so above it 4.5% of the time. Chapter 3 defines confidence intervals and shows what this does to them.



Simulation, 20,000 samples per size: means of skewed revenue against the normal curve the central limit theorem predicts. At $n = 5$ most samples contain no buyer; at $n = 30$ the distribution is still skewed; at $n = 1,000$ it is close to normal.

In ABMeter, every test and interval of the next chapter rests on this theorem: the metric value of a group of users is treated as approximately normal around its true value.

Exercises

1. One user's revenue has $\sigma = 22.98$ (the standard deviation of one user's value). What is the standard error of the mean of 100 users, and of 10,000 users?
2. One user's revenue has skewness $\gamma = 10.6$. What is the skewness of the mean of 30 users? How many users are needed for the skewness of the mean to fall to 0.1?
3. Use the law of large numbers bound with $\sigma = 23$ (the standard deviation of one user's value), $n = 10,000$ and $\varepsilon = 1$: at most how likely is the sample mean to be more than \$1 from μ (the mean of one user's value)? What does the central limit theorem give for the same probability?
4. Revenue per user has $\sigma = 20$, so by the central limit theorem $\sqrt{n} \cdot (\bar{X} - \mu)$ is approximately $N(0, 400)$. For $n = 400$, approximately how likely is $\bar{X}$ to be more than 2 away from μ (use $\Phi(2) = 0.9772$)? How many users are needed for the 95% band of $\bar{X} - \mu$ to be ± 0.5 ?

Hypothesis tests and confidence intervals

Two groups of users saw two versions of a page, and their conversion rates differ. A test asks whether a difference that large is plausible when the versions do not matter; a confidence interval says which true differences are consistent with the data. This chapter defines both precisely, builds the two-sample z-test, and shows that the test and the interval are two views of one calculation.

Tests and p-values

The idea

Picture a world in which the change did nothing, and the two groups differ only because different users happened to land in each. The p-value says how surprising your result would be in that world: the chance of a difference at least as large as the one you saw. A small p-value means “this would be rare if nothing had changed”. It does not say how likely it is that nothing changed.

Why test the hypothesis you do not believe

You want to know whether the change helped. The trouble is that “it helped” is not one claim but a family: a lift of 0.1 points, of 2 points, of 20. No single distribution of the data follows from it, so nothing can be computed. “It did nothing” is one claim, and it pins down everything: the two groups are then samples from the same population, and the difference of their averages has a known distribution centred at 0, with the standard error of Chapter 1. Fisher’s move, in the 1920s, was to compute in the one world where computation is possible and ask whether the data look ordinary there. If they do not, that world is doubted. Nothing in the test ever says how likely the change is to have helped; it only grades the data against the world with no effect.

Definition 3.1 (Hypotheses and test statistic)

The **null hypothesis** H_0 is the claim to test against, usually “the true difference is 0”. The **alternative** H_1 is its complement, here “the true difference is not 0”. A **test statistic** is a number computed from the data whose distribution under H_0 is known, at least approximately.

Example 3.2 (A new checkout button)

A team tests a new checkout button against the old one. H_0 : both buttons convert at the same rate. H_1 : the rates differ. The test statistic of the next section divides the difference of the two observed rates by its standard error; under H_0 it is approximately standard normal, so its distribution is known without knowing the conversion rate itself.

Definition 3.3 (p-value)

The **p-value** is the probability, computed under H_0 (written P_{H_0}), that the test statistic is at least as extreme as the value observed. For a statistic Z that is standard normal under H_0 and an observed value z , the two-sided p-value is, with Φ the standard normal distribution function,

$$\begin{aligned} p &= P_{H_0}(|Z| \geq |z|) && \text{(definition)} \\ &= P_{H_0}(Z \geq |z|) + P_{H_0}(Z \leq -|z|) && \text{(the two tails)} \\ &= 2(1 - \Phi(|z|)) && \text{(the normal is symmetric, so each tail is } 1 - \Phi(|z|)) \end{aligned}$$

Example 3.4 (An email subject line)

A new subject line gives a test statistic $z = 2.1$. With $\Phi(2.1) = 0.9821$:

$$\begin{aligned} p &= 2(1 - \Phi(2.1)) && \text{(two-sided p-value)} \\ &= 2 \times 0.0179 = 0.036 && \text{(table value)} \end{aligned}$$

If the subject line made no difference, a statistic at least this far from 0 would turn up in about 3.6% of experiments.

Example 3.5 (What $p = 0.04$ does not mean)

A team runs 1,000 experiments in a year. Suppose 100 of the changes really work, and the test detects a working change 80% of the time; the other 900 do nothing, and each of those still gives $p < 0.05$ with probability 0.05.

$$\begin{aligned} \text{significant and real} &= 100 \times 0.80 = 80 && \text{(detected working changes)} \\ \text{significant and null} &= 900 \times 0.05 = 45 && \text{(false positives)} \\ \text{share of significant results that are null} &= 45/(80 + 45) = 0.36 && \text{(of 125 significant results)} \end{aligned}$$

Every one of the 45 false positives had $p < 0.05$, yet more than a third of the “wins” did nothing. A p-value is computed in the world where H_0 is true; it is not the probability that H_0 is true, which depends on how many tested changes work at all.

Remark 3.6 (What a p-value is not)

- It is not the probability that H_0 is true. That would need a prior probability for H_0 , which the test does not use.
- It is not the probability that the result is a fluke, for the same reason.
- $1 - p$ is not the probability that the treatment works.
- It is not a measure of effect size: a tiny effect gives a tiny p-value when n is large.
- A large p-value does not show that H_0 is true. It shows that the data cannot tell.

Definition 3.7 (Significance level, errors and power)

A test at **level** α rejects H_0 when $p < \alpha$; the usual choice is $\alpha = 0.05$. A **type I error** is rejecting a true H_0 (a false positive). A **type II error** is not rejecting a false H_0 (a false negative). The **power** of a test is the probability of rejecting H_0 when a given alternative is true, $1 - \beta$, where β is the type II error rate.

Example 3.8 (The year of experiments, by error type)

In the previous example the level is $\alpha = 0.05$ (the test level), so 45 of the 900 null changes are type I errors. The power against the working changes is 0.80, so $\beta = 0.20$ (the type II error rate) and 20 of the 100 working changes are missed: type II errors.

Proposition 3.9 (p-values are uniform under the null)

If the test statistic has a continuous distribution under H_0 , then the p-value is uniformly distributed on $[0, 1]$ under H_0 : every subinterval of $[0, 1]$ has probability equal to its length. In particular $P_{H_0}(p < \alpha) = \alpha$: a level- α test makes a type I error with probability α .

Proof.

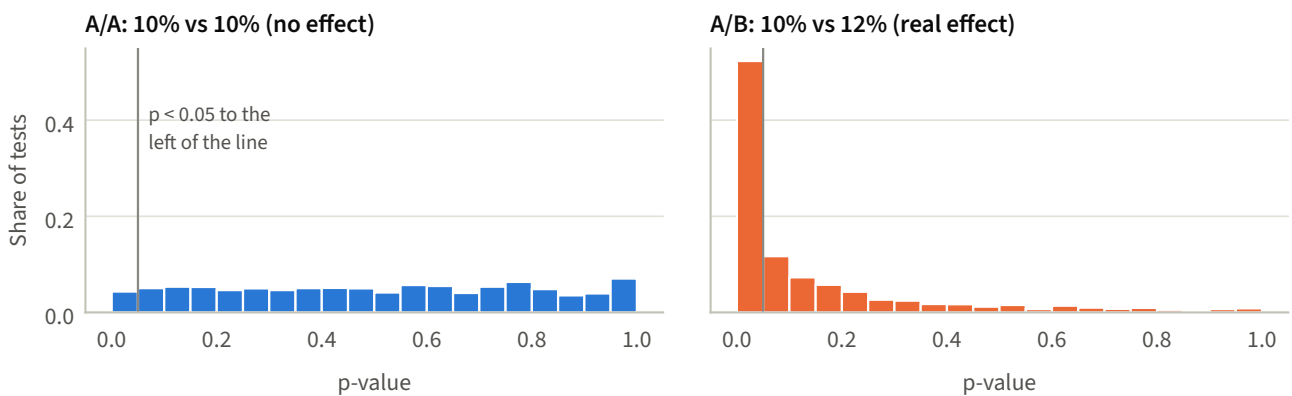
Let F be the distribution function of $|Z|$ under H_0 , so $p = 1 - F(|Z|)$, and let F be continuous. For u in $[0, 1]$:

$$\begin{aligned} P_{H_0}(F(|Z|) \leq u) &= P_{H_0}(|Z| \leq F^{-1}(u)) && (F \text{ is increasing}) \\ &= F(F^{-1}(u)) && (\text{definition of } F) \\ &= u && (F \text{ is continuous}) \end{aligned}$$

So $F(|Z|)$ is uniform on $[0, 1]$ (the probability integral transform), and so is $p = 1 - F(|Z|)$. ■

Example 3.10 (A/A and A/B simulations)

An **A/A test** shows both groups the same version, so H_0 is true; an **A/B test** shows them different versions. In 2,000 simulated A/A experiments, each with 2,000 users per group converting at 10%, 4.3% gave $p < 0.05$ (expected 5%, with standard error 0.5 points over 2,000 experiments, the square root of $0.05 \times 0.95 / 2000$), and each tenth of $[0, 1]$ received between 8% and 12% of the p-values (expected 10% each, standard error 0.7 points). In 2,000 A/B experiments with the treatment converting at 12%, 52.3% gave $p < 0.05$ and 64% of the p-values fell below 0.1. The 52.3% is the power of this design; the power formula of Chapter 8 gives 52.5% for it, from a standard error of 0.0099, the square root of $(0.09 + 0.1056) / 2000$, and an effect of $0.02 / 0.0099 = 2.02$ standard errors.



Simulation, 2,000 experiments each. Without an effect, p-values spread evenly over $[0, 1]$, so 5% fall below 0.05 by chance. With an effect they pile up near 0; here only about half fall below 0.05, because 2,000 users per group is too few for a 2-point difference.

The A/A share agrees with 5% within its sampling error, and the A/B share is the power of this design against a 2-point effect: only about half of such experiments find it.

The two-sample z-test

The idea

The two groups' averages always differ a little, even when the change does nothing. The z-statistic measures the observed difference in units of its own noise: how many standard errors apart the two averages are. Two or more standard errors apart is unusual when nothing changed.

Proposition 3.11 (Variance of a difference of means)

Let a treatment group T and a control group C be independent i.i.d. samples of sizes n_T and n_C , with per-user variances σ_T^2 and σ_C^2 and sample means $\bar{X}_T$ and $\bar{X}_C$. Then

$$\text{Var}(\bar{X}_T - \bar{X}_C) = \frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}$$

Proof.

$$\begin{aligned}\text{Var}(\bar{X}_T - \bar{X}_C) &= \text{Var}(\bar{X}_T) + \text{Var}(-\bar{X}_C) + 2 \text{Cov}(\bar{X}_T, -\bar{X}_C) && \text{(Proposition 1.17)} \\ &= \text{Var}(\bar{X}_T) + \text{Var}(\bar{X}_C) + 2 \text{Cov}(\bar{X}_T, -\bar{X}_C) && \text{(Proposition 1.5 with } a = -1\text{)} \\ &= \text{Var}(\bar{X}_T) + \text{Var}(\bar{X}_C) && \text{(independent groups: covariance 0)} \\ &= \sigma_T^2/n_T + \sigma_C^2/n_C && \text{(Proposition 1.21 in each group)}\end{aligned}$$

Example 3.12 (Noise in a signup test)

A signup form converts at 10% in control and 12% in treatment, with 1,000 users in each group. The 0/1 variances are $0.1 \times 0.9 = 0.09$ and $0.12 \times 0.88 = 0.1056$.

$$\begin{aligned}\text{Var}(\bar{X}_T - \bar{X}_C) &= 0.1056/1000 + 0.09/1000 && \text{(the proposition)} \\ &= 0.0001956 && \text{(add)} \\ \text{its standard deviation} &= \sqrt{0.0001956} = 0.014 && \text{(square root)}\end{aligned}$$

The true difference, 0.02, is only 1.4 times the noise of the observed difference, so this design will often miss it.

Replacing each per-user variance by its sample estimate gives the standard error of the difference.

Why divide by the standard error

A difference of \$2 in revenue per user and a difference of 2 points in conversion cannot be compared, and neither can a \$2 difference over 50 users and one over 50,000. Dividing a difference by its own standard error converts it into a unit that every metric and every sample size share: how many standard errors of noise it spans. Under H_0 the result is approximately standard normal by the central limit theorem, so one table serves every test, and 1.96 means the same thing for a conversion rate as for revenue. The z-statistic is the difference, read in units of its own noise.

Definition 3.13 (Two-sample z-test with unequal variances)

With the groups' sample variances s_T^2 and s_C^2 , write $SE_T^2 = s_T^2/n_T$ and $SE_C^2 = s_C^2/n_C$. The test statistic for H_0 : no difference is

$$z = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{SE_T^2 + SE_C^2}}$$

The denominator is the standard error of the difference: the square root of the variance in the proposition above, with each σ^2 replaced by s^2 . By the central limit theorem z is approximately standard normal under H_0 when both groups are large.

Example 3.14 (Ten users per group, by hand)

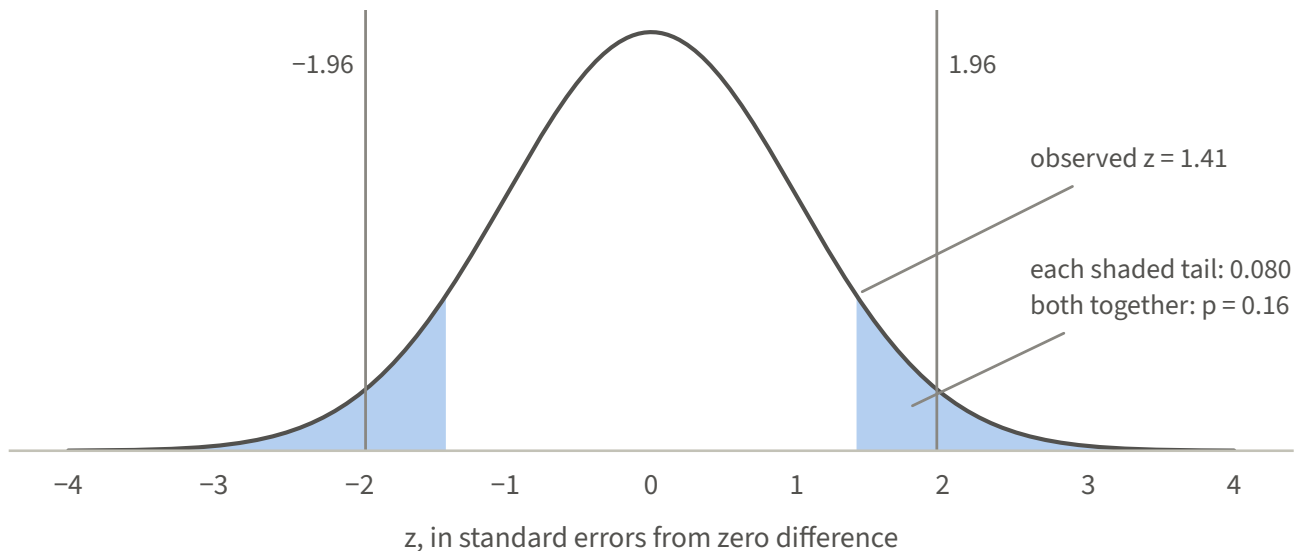
Control: 2 of 10 users converted, so the observed rate is $\hat{p}_C = 0.2$. Treatment: 5 of 10, $\hat{p}_T = 0.5$. With $n = 10$ users per group:

$$\begin{aligned} s_C^2 &= \frac{10}{9} \times 0.2 \times 0.8 = 0.17778 & (0/1 \text{ shortcut } s^2 = n \hat{p}(1 - \hat{p})/(n - 1)) \\ SE_C^2 &= 0.17778/10 = 0.017778 & (SE^2 = s^2/n) \\ s_T^2 &= \frac{10}{9} \times 0.5 \times 0.5 = 0.27778 & (0/1 \text{ shortcut}) \\ SE_T^2 &= 0.27778/10 = 0.027778 & (SE^2 = s^2/n) \\ SE &= \sqrt{0.017778 + 0.027778} = 0.21344 & (\text{standard error of the difference}) \\ z &= (0.5 - 0.2)/0.21344 = 1.4056 & (\text{definition of } z) \\ p &= 2(1 - \Phi(1.4056)) = 0.1599 & (\text{two-sided p-value}) \end{aligned}$$

A difference of 30 points is not significant with 10 users per group. The same rates with more users:

$$\begin{aligned} n = 100: \quad SE &= \sqrt{\frac{100}{99} (0.16 + 0.25)/100} = 0.0644 & (0/1 \text{ shortcut in each group, } SE^2 = s^2/n) \\ z &= 0.3/0.0644 = 4.66 & (\text{definition of } z) \\ p &= 2(1 - \Phi(4.66)) = 3.1 \times 10^{-6} & (\text{two-sided p-value}) \\ n = 1000: \quad SE &= \sqrt{\frac{1000}{999} (0.41)/1000} = 0.0203 & (\text{as above}) \\ z &= 0.3/0.0203 = 14.81 & (\text{definition of } z) \end{aligned}$$

The effect stays the same; the standard error shrinks like $1/\sqrt{n}$. At $z = 14.81$ a computer prints the p-value as 0.0: once z exceeds about 8.3, $1 - \Phi(z)$ is below 5×10^{-17} and $\Phi(z)$ rounds to 1 in double precision.



The standard normal density: the distribution of z if the two versions do not differ. The observed $z = 1.41$ leaves 0.080 in each tail, so $p = 0.16$. A level-0.05 test rejects when z falls outside ± 1.96 .

Remark 3.15 (Relatives of the z -test)

- **Welch's t-test** uses the same statistic and compares it with a t -distribution, a bell with heavier tails than the normal, whose shape parameter, the degrees of freedom, is estimated from the data. For large groups the t -distribution is the normal one.
- **Student's t-test** and the **pooled two-proportion test** assume both groups have the same variance and estimate it from both together. They agree with the unequal-variance test when the groups are alike and differ when sizes and variances differ.

With 3,000 control users (300 buyers, rate 0.10) and 1,000 treatment users (125 buyers, rate 0.125):

$$\begin{aligned}
 SE^2 &= \frac{3000}{2999} \frac{0.1 \times 0.9}{3000} + \frac{1000}{999} \frac{0.125 \times 0.875}{1000} = 0.0001395 && \text{(0/1 shortcut in each group)} \\
 z &= 0.025 / \sqrt{0.0001395} = 2.12 && \text{(definition of } z; p = 0.034) \\
 \bar{p} &= 425/4000 = 0.10625 && \text{(pooled rate of both groups)} \\
 z_{\text{pooled}} &= \frac{0.025}{\sqrt{0.10625 \times 0.89375 (1/3000 + 1/1000)}} = 2.22 && \text{(pooled variance; } p = 0.026)
 \end{aligned}$$

The pooled test gives the smaller group the larger group's variance, which is too small here. With 10 users per group (2 against 5 buyers) the normal tail gives $p = 0.160$ ([Example 3.14](#)), while a t -table with the 17.2 degrees of freedom that Welch's formula assigns to these two samples gives 0.178: the z -test is optimistic in small samples because it ignores the uncertainty in s^2 (the sample variance).

Confidence intervals

The idea

A confidence interval is the range of true differences that the data cannot rule out. The 95% is a property of the procedure, not of one interval: if you re-ran the experiment many times and built the interval each time, about 95 of every 100 intervals would contain the true difference. Any single interval either contains it or does not.

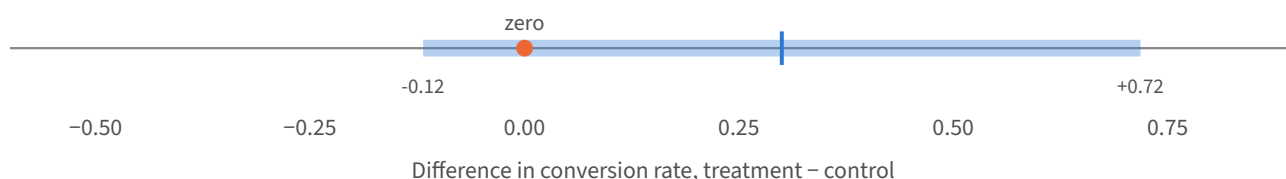
Why an interval, and why 1.96

A test answers one question: whether 0 is plausible. The same calculation can be run for any other candidate: a true difference d is plausible when the observed difference lies within 1.96 standard errors of it. Collecting every plausible d gives an interval, and because the condition is symmetric, that interval is the estimate plus or minus 1.96 standard errors. The 1.96 is the width that leaves 2.5% in each tail of the normal curve and 95% in the middle; a 90% interval uses 1.645 and a 99% interval 2.576. Neyman proposed the construction in 1937 because a p-value alone says nothing about how large the effect could be.

If the versions did not differ: ordinary observed differences



True differences the data cannot rule out: the 95% interval



The ten-user test of [Example 3.14](#) on two number lines. Top: the differences that would be ordinary if the versions did not differ, a band of ± 1.96 SE around 0, with the observed 0.3 inside it. Bottom: the true differences the data cannot rule out, the same band moved to sit around 0.3, with 0 inside it. The two bands have the same width, so one contains the observed difference exactly when the other contains 0.

Definition 3.16 (Confidence interval and coverage)

A **95% confidence interval** for an unknown quantity θ is an interval computed from the data that contains θ with probability 0.95 over repeated samples. That probability is its **coverage**. For the difference of means the normal approximation gives, with SE the standard error of the difference and $1.96 = z_{0.975}$ (the standard normal quantile),

$$(\bar{X}_T - \bar{X}_C) \pm 1.96 \text{ SE}$$

Example 3.17 (The ten-user interval)

In [Example 3.14](#) the observed difference is 0.3 and $\text{SE} = 0.21344$ (the standard error of the difference).

$$\begin{aligned} 1.96 \times 0.21344 &= 0.4183 && \text{(half-width)} \\ 0.3 \pm 0.4183 &= [-0.118, 0.718] && \text{(the interval)} \end{aligned}$$

Example 3.18 (What the interval does not say)

The interval $[-0.118, 0.718]$ contains 0, and it is tempting to read “no effect”. It also contains 0.7, a 70-point lift. The data are consistent with no effect and with a huge one; they cannot tell the two apart. Nor is there a 95% probability that the true difference lies in $[-0.118, 0.718]$: the true difference is a fixed number, and this interval either contains it or does not. The 95% describes how often the procedure succeeds across re-runs.

Proposition 3.19 (Duality of tests and intervals)

The 95% interval for the difference excludes 0 if and only if the two-sided z-test rejects at level 0.05.

Proof.

$$\begin{aligned}
 0 \notin (\bar{X}_T - \bar{X}_C) \pm 1.96 \text{ SE} &\iff |\bar{X}_T - \bar{X}_C| > 1.96 \text{ SE} && (0 \text{ lies outside the interval}) \\
 &\iff |z| > 1.96 && (\text{divide by SE} > 0) \\
 &\iff 2(1 - \Phi(|z|)) < 2(1 - \Phi(1.96)) && (\Phi \text{ is increasing}) \\
 &\iff p < 0.05 && (2(1 - \Phi(1.96)) = 0.05)
 \end{aligned}$$

Example 3.20 (Ten and a hundred users)

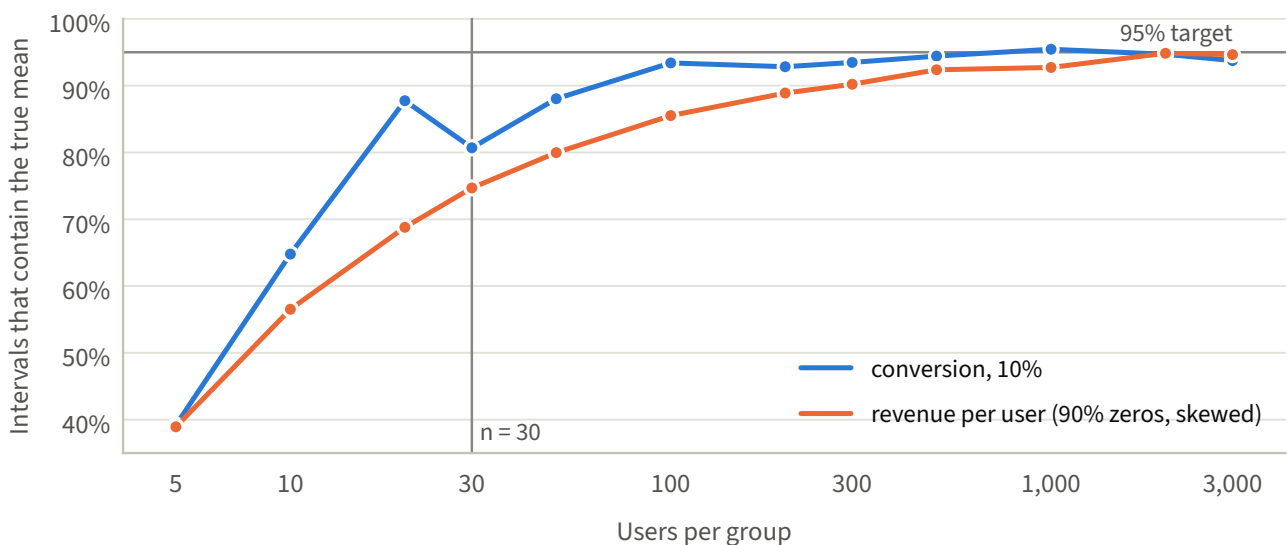
With 10 users per group the interval $[-0.118, 0.718]$ contains 0 and $p = 0.16$ (the p-value) is above 0.05. With 100 users per group and the same rates, $\text{SE} = 0.0644$ from the z-test example, so the interval is $0.3 \pm 1.96 \times 0.0644 = [0.174, 0.426]$, which excludes 0, and $p = 3.1 \times 10^{-6}$ is below 0.05. The two verdicts always agree.

Example 3.21 (How fast coverage reaches 95%)

The interval $\text{mean} \pm 1.96 \text{ SE}$ relies on the central limit theorem, so its real coverage at small n is below 95%. A simulation of 20,000 samples of each size, for a 10% conversion rate and for the skewed revenue of Chapter 2, counts how often the interval contains the true mean:

users	10% conversion	skewed revenue
5	39.2%	38.9%
30	81.0%	74.7%
100	93.0%	85.8%
1,000	95.2%	93.3%
3,000	95.0%	94.5%

A “95%” interval from 30 users is really an 81% interval for a 10% conversion and a 75% interval for skewed revenue, and the matching p-values are too small. The common rule of “30 is enough” holds only for nearly symmetric data. At five users many samples contain no conversion at all; their standard error is 0 and the interval cannot contain the truth.



Simulation, 20,000 samples per size: real coverage of the nominal 95% interval. At 30 users per group it is 81% for a 10% conversion rate and 75% for skewed revenue; skewed revenue needs about 1,000 users to come close to 95%.

In ABMeter, the p-value and the absolute confidence interval of each treatment group are this z-test and this interval, computed against the control group.

Exercises

1. Control converts 100 of 1,000 users and the treatment 125 of 1,000. Compute the standard error of the difference, z (the test statistic), the two-sided p-value (use $\Phi(1.77) = 0.9616$) and the 95% interval. Is the difference significant at level 0.05?
2. A test gives $p = 0.03$. Which of these statements are true? (a) The probability that H_0 is true is 3%. (b) If H_0 were true, a statistic at least this extreme would occur with probability 0.03. (c) The 95% interval for the difference excludes 0. (d) The effect is large.
3. You run 20 independent A/A tests at level 0.05. What is the probability that at least one of them is significant?

Randomized experiments

The z-test of Chapter 3 says whether two groups differ by more than chance. It does not say why. This chapter shows why random assignment lets you read the difference as the effect of the treatment, using the potential-outcomes model, and how a hash function provides random assignment that is repeatable for each user.

Potential outcomes

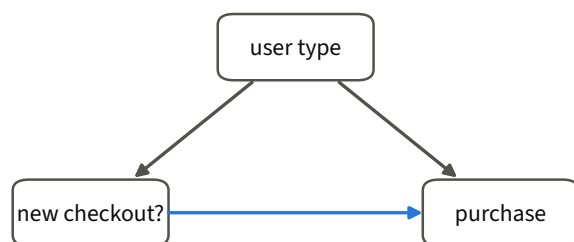
The idea

Think of every user as carrying two answers: what they would do with the old version and what they would do with the new one. You only ever see one of the two. If a coin decides who gets which version, the treated users are a fair sample of everyone, and so are the control users, so the two group averages estimate the averages of the two answers over all users.

Seeing and doing

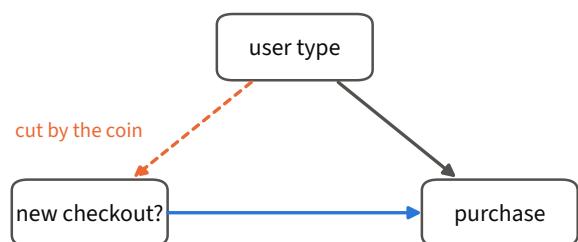
Two questions sound alike and are not. “Among the users who chose the new checkout, how many bought?” observes a group that chose itself. “If we gave a user the new checkout, how many would buy?” asks what the checkout does. Pearl (2009) writes the first as $P(\text{purchase} \mid \text{checkout})$, seeing, and the second as $P(\text{purchase} \mid \text{do}(\text{checkout}))$, doing, and this chapter is about the gap between them. In the diagram, user type points at both the choice of checkout and the purchase: heavy users opt in more and buy more. A comparison of opt-ins with the rest therefore picks up the type’s effect along with the checkout’s, through the route that runs back through type. A coin flip removes the arrow from type into the choice, since the coin does not know the user’s type, and the only route left from the group to the purchase is the checkout itself. That is what “do” means in practice, and it is why the rest of the chapter can treat the difference of two group averages as the effect. Everything below is written in the language of potential outcomes, which gives the same answer for a randomized test and carries the standard error with it.

Users choose the checkout



two routes from the choice to the purchase

A coin chooses the checkout



one route: the checkout itself

Left: when users choose, user type drives both the choice and the purchase, so the choice and the purchase are linked by two routes, and a comparison of the two groups mixes them. Right: a coin decides the group, which cuts the arrow from type into the choice; the one route left is the checkout's own effect.

Definition 4.1 (Potential outcomes and the average treatment effect)

Each user i has two **potential outcomes**: $Y_i(1)$, the outcome if treated, and $Y_i(0)$, the outcome if not. The **individual effect** is $Y_i(1) - Y_i(0)$, and the **average treatment effect** (ATE) over n users is

$$\text{ATE} = \frac{1}{n} \sum_{i=1}^n (Y_i(1) - Y_i(0))$$

Example 4.2 (Six users and a new checkout)

Six users would make these numbers of purchases without and with a new checkout:

User	1	2	3	4	5	6
$Y_i(0)$	0	1	0	2	0	1
$Y_i(1)$	1	1	0	3	1	1

The individual effects $Y_i(1) - Y_i(0)$ are 1, 0, 0, 1, 1, 0, so the ATE (the average treatment effect) is $3/6 = 0.5$. In a real test each user shows only one of the two rows.

Only one of the two outcomes is ever observed for a user: the one for the version the user saw. The individual effect is never observed. The average can still be estimated, if the choice of who is treated is random.

Proposition 4.3 (Randomization gives an unbiased estimate)

Assign n_T of the n users to treatment completely at random, every subset of size n_T being equally likely, and the rest to control. Let $\bar{Y}_T$ be the mean of $Y_i(1)$ over the treated users and $\bar{Y}_C$ the mean of $Y_i(0)$ over the control users, the outcomes actually observed. Then the difference of group means is unbiased for the ATE:

$$\mathbb{E}[\bar{Y}_T - \bar{Y}_C] = \text{ATE}$$

Proof.

Let $T_i = 1$ if user i is treated and 0 otherwise, so that $\bar{Y}_T = \sum_i T_i Y_i(1)/n_T$. Each user is treated with probability n_T/n , so $\mathbb{E}[T_i] = n_T/n$. Only the assignment T_i is random; the potential outcomes are fixed numbers.

$$\begin{aligned} \mathbb{E}[\bar{Y}_T] &= \frac{1}{n_T} \sum_{i=1}^n \mathbb{E}[T_i] Y_i(1) && \text{(linearity)} \\ &= \frac{1}{n_T} \sum_{i=1}^n \frac{n_T}{n} Y_i(1) && (\mathbb{E}[T_i] = n_T/n) \\ &= \frac{1}{n} \sum_{i=1}^n Y_i(1) && \text{(cancel } n_T) \end{aligned}$$

In the same way $\mathbb{E}[\bar{Y}_C]$ is the average of all $Y_i(0)$, so

$$\begin{aligned} \mathbb{E}[\bar{Y}_T - \bar{Y}_C] &= \frac{1}{n} \sum_{i=1}^n Y_i(1) - \frac{1}{n} \sum_{i=1}^n Y_i(0) && \text{(linearity)} \\ &= \frac{1}{n} \sum_{i=1}^n (Y_i(1) - Y_i(0)) = \text{ATE} && \text{(definition of the ATE)} \end{aligned}$$

■

The only randomness here is the assignment. Users and their potential outcomes are fixed; the proof needs no model for how users behave.

Example 4.4 (Every assignment of the six users)

Treat 3 of the users of [Example 4.2](#). There are $(6 \times 5 \times 4)/(3 \times 2 \times 1) = 20$ ways to choose them. Treating users 1, 2 and 3 reveals their $Y_i(1)$ and the $Y_i(0)$ of users 4, 5 and 6:

$$\begin{aligned}\bar{Y}_T &= (1 + 1 + 0)/3 = 2/3 && \text{(treated users 1, 2, 3 show } Y_i(1)) \\ \bar{Y}_C &= (2 + 0 + 1)/3 = 1 && \text{(control users 4, 5, 6 show } Y_i(0)) \\ \bar{Y}_T - \bar{Y}_C &= -1/3 && \text{(subtract)}\end{aligned}$$

One assignment can be far off. The lowest estimate treats users 1, 3 and 5, whose $Y_i(1)$ are 1, 0, 1, and leaves users 2, 4 and 6, whose $Y_i(0)$ are 1, 2, 1, in control: $2/3 - 4/3 = -2/3$. The highest does the opposite, treating users 2, 4 and 6 ($Y_i(1) = 1, 3, 1$) against control users 1, 3 and 5 ($Y_i(0) = 0, 0, 0$): $5/3 - 0 = 5/3$. Averaged over all 20 assignments the estimate is exactly 0.5, the ATE.

When users choose their own version, the proof fails: the probability of being treated depends on the user, and if it is related to $Y_i(0)$ the estimate is biased. This is **selection bias**.

Example 4.5 (A new checkout that users can choose)

A site has 20,000 users, of whom 30% are heavy users: 6,000 heavy and 14,000 light. A heavy user buys with probability 0.30, a light user with probability 0.05. A new checkout raises both probabilities by 0.02, to 0.32 and 0.07, so the true effect is +0.020 for everyone and the ATE (the average treatment effect) is +0.020. All probabilities below are written as fractions. The simulation draws each user's type, choice and purchase at random with these probabilities; the figures marked "simulated" are one such draw of the 20,000 users.

Naive: compare users who chose the new checkout with those who did not

Suppose users choose: "opting in" means choosing the new checkout, and 70% of heavy users opt in but only 20% of light users. The expected group sizes are

$$\begin{aligned}\text{treated} &= 0.70 \times 6000 + 0.20 \times 14000 = 4200 + 2800 = 7000 && \text{(heavy plus light opt-ins)} \\ \text{untreated} &= 1800 + 11200 = 13000 && \text{(the rest)}\end{aligned}$$

so the treated group is 60% heavy (4200 of 7000) and the untreated group is 14% heavy (1800 of 13000). Each group's purchase rate is a weighted average of the two types' rates:

$$\begin{aligned}\text{treated rate} &= (4200 \times 0.32 + 2800 \times 0.07)/7000 = 0.220 && \text{(new-checkout rates)} \\ \text{untreated rate} &= (1800 \times 0.30 + 11200 \times 0.05)/13000 = 0.0846 && \text{(old-checkout rates)} \\ \text{difference} &= 0.220 - 0.0846 = 0.135 && \text{(expected naive estimate)}\end{aligned}$$

Suppose for a moment that the new checkout did nothing, so that everyone buys at the old-checkout rates of 0.30 and 0.05. The treated group is still 60% heavy and 40% light, so its rate would be $0.60 \times 0.30 + 0.40 \times 0.05 = 0.200$, while the untreated group, which already buys at the old rates, stays at 0.0846. The gap of $0.200 - 0.0846 = 0.115$ therefore comes from composition alone: the treated group simply contains more heavy users. The true effect adds the remaining $0.220 - 0.200 = 0.020$. In the simulated draw, 4,175 heavy and 2,790 light users opted in, with purchase rate 0.2263, against 0.0822 for the 1,830 heavy and 11,205 light users who did not, so the estimate is $0.2263 - 0.0822 = +0.144$, seven times the true +0.020. Its standard error, by the formula of Chapter 3 with the 0/1 variance $\hat{p}(1 - \hat{p})$ in each group, is

$$SE = \sqrt{\frac{0.2263 \times 0.7737}{6965} + \frac{0.0822 \times 0.9178}{13035}} = 0.0056$$

so the 95% interval is $0.144 \pm 1.96 \times 0.0056 = [+0.133, +0.155]$: narrow, and nowhere near the truth.

Correct: assign at random

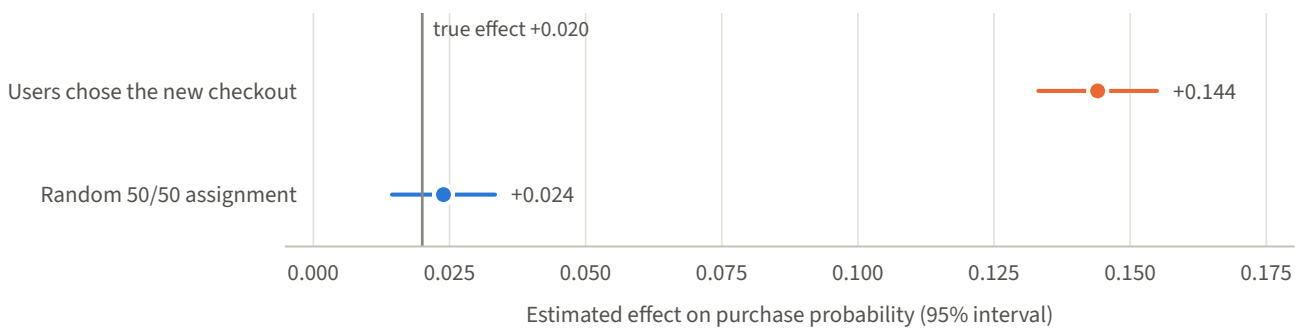
A coin flip gives every user probability $\frac{1}{2}$ of the new checkout, whatever their type. The realised share of heavy users in treatment is a random variable: with 6,000 heavy users its standard deviation is about 0.0065, the square root of $0.25 / 6000$, so the simulated 50.7% (3,046 of 6,005) is an ordinary draw, about one standard deviation above $\frac{1}{2}$; 50.2% of light users (7,032 of 13,995) landed in treatment. Randomization balances the groups approximately, not exactly, and the standard error accounts for the remainder. The expected rates are

$$\begin{aligned}\text{treated rate} &= 0.30 \times 0.32 + 0.70 \times 0.07 = 0.145 && \text{(both types in their population shares)} \\ \text{untreated rate} &= 0.30 \times 0.30 + 0.70 \times 0.05 = 0.125 && \text{(the same shares, old checkout)} \\ \text{difference} &= 0.145 - 0.125 = 0.020 && \text{(the true effect)}\end{aligned}$$

The simulated groups have 10,078 and 9,922 users, with purchase rates 0.1467 and 0.1228, so the estimate is $0.1467 - 0.1228 = +0.024$, and

$$\text{SE} = \sqrt{\frac{0.1467 \times 0.8533}{10078} + \frac{0.1228 \times 0.8772}{9922}} = 0.0048$$

gives the 95% interval $0.024 \pm 1.96 \times 0.0048 = [+0.014, +0.033]$, which contains +0.020.



Simulation, one draw of 20,000 users with the probabilities of the example. The same users and the same true effect (+0.020): opt-in mixes the effect with who opted in; random assignment recovers it within its 95% interval.

Assignment by hashing

The idea

An online test needs a coin that gives each user the same side every time they come back, on every server, without storing a table. A hash function is such a coin: it scrambles the user id into a number that looks random across users but is always the same for the same user. Cutting the range of that number into 100 equal slices gives each user a fixed bucket.

Definition 4.6 (Hash bucket)

Let h be a cryptographic hash function such as SHA-256, and let a **salt** s be a fixed string chosen for one use of the hash. Let $\text{num}(s, u)$ be the first 64 bits of h applied to the string “salt:user” for user id u , read as an integer in $[0, 2^{64})$. With $\lfloor x \rfloor$ the floor of x (the largest integer at most x), the **bucket** of u under s is

$$b(s, u) = \left\lfloor \frac{100 \cdot \text{num}(s, u)}{2^{64}} \right\rfloor + 1$$

A group is a range of buckets: with a 50/50 split, buckets 1 to 50 are control and 51 to 100 are treatment.

Example 4.7 (One user's bucket)

Suppose the hash of “pricing-test:u42” gives $\text{num}/2^{64} = 0.3721$, so the number sits 37.21% of the way through its range.

$$\begin{aligned} b &= \lfloor 100 \times 0.3721 \rfloor + 1 && \text{(definition of the bucket)} \\ &= \lfloor 37.21 \rfloor + 1 = 38 && \text{(round down, add 1)} \end{aligned}$$

Bucket 38 is in 1 to 50, so user u42 sees the control version, today and on every later visit. Computers evaluate the floor as $(\text{num} \times 100)$ shifted right by 64 bits, which needs no division; Lemire (2018) analyses this multiply-and-shift mapping.

Proposition 4.8 (Buckets are uniform)

If $\text{num}(s, u)$ were uniform on $[0, 2^{64})$, each bucket would receive either $\lfloor 2^{64}/100 \rfloor$ or $\lceil 2^{64}/100 \rceil$ of the possible values, where $\lceil x \rceil$ is the ceiling (the smallest integer at least x), so every bucket has probability within 2^{-64} of $1/100$.

Proof.

Bucket k receives the integers x with $(k-1)2^{64}/100 \leq x < k2^{64}/100$, a half-open interval of length $2^{64}/100$. Such an interval contains either the floor or the ceiling of its length in integers. ■

Example 4.9 (Sixteen values in three buckets)

Shrink the range to 16 values, 0 to 15, and use 3 buckets: $b = \lfloor 3x/16 \rfloor + 1$. Bucket 1 receives $x = 0, \dots, 5$ (6 values, since $3 \times 5/16 = 0.94$), bucket 2 receives $x = 6, \dots, 10$ (5 values) and bucket 3 receives $x = 11, \dots, 15$ (5 values). The ideal share is $16/3 = 5.33$ values, and each bucket gets its floor 5 or its ceiling 6.

A cryptographic hash is not random, but no practical test can tell its output on distinct inputs from uniform random bits. Treating different salts as independent random functions is a modeling assumption, and hashing 100,000 ids (“user-0” to “user-99999”) under two salts checks its consequences. Under one salt, each of the 100 buckets expects $100,000 / 100 = 1,000$ ids, with standard deviation about 31.5, the square root of $100,000 \times 0.01 \times 0.99$ (Example 1.9 summed over the ids); the observed counts range from 916 to 1,090, that is from 2.7 standard deviations below to 2.9 above, as the extremes of 100 such counts should. A chi-square test, which adds up the squared gaps between the 100 counts and 1,000 in units of their variance, gives 71.3 on 99 degrees of freedom, $p = 0.98$: no evidence against uniform buckets. Under two salts, 24.7% of ids have both buckets at most 50, where independence predicts $0.5 \times 0.5 = 25\%$ with standard error 0.14 points, the square root of $0.25 \times 0.75 / 100,000$, and the correlation of the two buckets is -0.002 .

Example 4.10 (Twenty users, two steps)

A pool of users is shared by several tests; this test uses buckets 1 to 50 of the pool salt. Inside the test, a second bucket under the test's own salt picks the group: 1 to 50 control, 51 to 100 treatment. The buckets below are SHA-256 outputs for the ids u01 to u20 and cannot be predicted by hand; what can be checked is the rule applied to them. Of the 20 users, these eight have a pool bucket of at most 50:

user	u02	u04	u06	u07	u09	u12	u19	u20
pool bucket	38	11	48	42	25	3	17	27
group bucket	53	9	58	66	55	60	94	40
group	T	C	T	T	T	T	T	C

The other twelve, such as u01 with pool bucket 75, are not in the test. Two of the eight are in control and six in treatment. Small samples of a fair split are lumpy; with 10,000 users the split is close to even, as the next example shows.

Why two salts? The second step must be independent of the first.

Naive: one salt for both steps

A user is in the test only if the pool bucket is at most 50. With the same salt, the group bucket is the same number, so every user in the test lands in control. Hashing 10,000 ids: 4,993 are in the test (expected $10,000 \times 0.5 = 5,000$, standard deviation 50, the square root of $10,000 \times 0.25$), all 4,993 in control and none in treatment.

Correct: a second, independent salt

Users in the test spread evenly over the group buckets: 2,522 in control and 2,471 in treatment, against an expected $4,993 / 2 \approx 2,497$ each with standard deviation about 35, the square root of $4,993 \times 0.25$.

Remark 4.11 (Ranges must not change during a test)

A bucket never changes, so a user's group is fixed as long as the ranges are. Changing a running 50/50 split to 40/60 moves buckets 41 to 50 from control to treatment. Those users saw both versions, and their outcomes no longer belong to one group.

In ABMeter, a space is the pool: its experiments take disjoint bucket ranges under the space salt, so a user is in at most one experiment per space, and each experiment's own salt picks the audience.

Exercises

- Four users have $Y(0) = (2, 0, 1, 1)$ and $Y(1) = (3, 0, 1, 2)$. Compute the ATE (the average treatment effect), and the estimate when users 1 and 3 are treated. Over all 6 ways to treat two users, what is the average estimate?
- A test splits its users 50/50 with control on buckets 1 to 50. The split is changed to 40/60 while it runs. What share of the test's users switch group?
- Test A uses pool buckets 1 to 50 under salt s_A , and test B uses buckets 1 to 30 under an independent salt s_B . What share of all users is in both tests? What share would be in both if the two tests used the same salt?

Ratio metrics and the unit of analysis

Revenue per user is a mean over users. Average order value is not: it divides one total by another. This chapter shows that the two obvious ways to compute a ratio disagree, says which one estimates what, and explains why the user, not the order, must be the unit of the statistics.

Mean of ratios and ratio of means

The idea

There are two natural ways to compute average order value. You can work out each user's own average order and then average those over users, or you can divide all the revenue by all the orders. The first treats every user as one vote, however many orders they placed; the second treats every order as one vote. Only the second is "the value of an order".

Definition 5.1 (Two ratio estimators)

Let user i have X_i (for example revenue) and $Y_i \geq 0$ (for example orders). The **mean of ratios** averages each user's own ratio over the m users with $Y_i > 0$:

$$M = \frac{1}{m} \sum_{i: Y_i > 0} \frac{X_i}{Y_i}$$

The **ratio of means** divides the two sample means over all n users:

$$R = \frac{\bar{X}}{\bar{Y}} = \frac{\sum_i X_i}{\sum_i Y_i}$$

The factor $1/n$ cancels in R , so the ratio of means is a ratio of totals.

Example 5.2 (Alice, Bob and eight users who bought nothing)

Ten users. Alice made one purchase of \$100. Bob made nine purchases of \$10 each. Eight others bought nothing.

Naive: mean of ratios

Alice's ratio is $100 / 1 = 100$, Bob's is $90 / 9 = 10$, and the eight others have $0/0$, which is undefined and dropped. $M = (100 + 10)/2 = 55$, from 2 of the 10 users.

Correct: ratio of means

$\bar{X} = 190/10 = 19$ and $\bar{Y} = 10/10 = 1$, so $R = 19$: total revenue over total orders, from all 10 users.

\$55 is not the value of an order (ten orders average \$19) and not the value of a user (users bring \$19 each). It weighs Alice's one order as much as Bob's nine.

Proposition 5.3 (What each estimator estimates)

Let the pairs (X_i, Y_i) be i.i.d. with means $\mu_X = \mathbb{E}[X_i]$ and $\mu_Y = \mathbb{E}[Y_i] \neq 0$. (a) $R = \bar{X}/\bar{Y}$ (the ratio of means) is a consistent estimator of μ_X/μ_Y , the ratio of the population means. (b) Adding users with $X_i = Y_i = 0$ does not change R . (c) M (the mean of ratios) is a mean of order values in which each user's orders together get the same weight, whatever their number.

Proof.

(a) By the law of large numbers $\bar{X} \rightarrow \mu_X$ and $\bar{Y} \rightarrow \mu_Y$, and since x/y is continuous where $y \neq 0$, $\bar{X}/\bar{Y} \rightarrow \mu_X/\mu_Y$ (the continuous mapping theorem). (b) Such users add 0 to both totals. (c) X_i/Y_i is user i 's average order value, and M averages these with equal weights $1/m$. ■

Example 5.4 (Doubling the non-buyers)

Add ten more users who bought nothing to [Example 5.2](#). Now $n = 20$:

$$\bar{X} = 190/20 = 9.5 \quad (\text{total revenue over 20 users})$$

$$\bar{Y} = 10/20 = 0.5 \quad (\text{total orders over 20 users})$$

$$R = 9.5/0.5 = 19 \quad (\text{the ratio of means})$$

Both means halve and R stays at 19, so two groups of different sizes can be compared on R directly. M also stays at 55, but for a different reason: it ignores non-buyers altogether.

The unit of analysis

The idea

The coin in an experiment is flipped per user, so users are the independent units. Bob's nine orders did not arrive as nine separate coin flips: they came as one package with Bob, and moved into one group together. Statistics that count each order as an independent observation pretend there were more coin flips than there were.

Definition 5.5 (Unit of randomization and unit of analysis)

The **unit of randomization** is what the assignment chooses: here, the user. The **unit of analysis** is what a metric averages over: users for revenue per user, orders for average order value. When the two differ, the observations of the analysis unit are not independent, and their sample variance does not give the standard error.

Example 5.6 (Ten orders that are really two users)

Take the ten orders of [Example 5.2](#) as if they were ten independent observations: one order of \$100 and nine of \$10. Their mean is 19, and the deviations are 81 and nine times -9 :

$$s^2 = (81^2 + 9 \times 9^2)/9 = (6561 + 729)/9 = 810 \quad (\text{sample variance of the ten orders})$$

$$s/\sqrt{n} = \sqrt{810}/\sqrt{10} = 28.46/3.16 = 9.00 \quad (\text{the order-level "standard error"})$$

Chapter 6 computes the correct standard error of R (the ratio of means) with users as the unit: 12.07. The order-level formula is a quarter too small, because nine of the ten "observations" are one person.

The fix is to keep the user as the observation. Per user, record the two numbers X_i and Y_i ; these pairs are independent across users. Then R is a smooth function of two user-level means, and Chapter 6 computes its standard error with the delta method.

A ratio can move for the wrong reason

The idea

A ratio has an outcome in its denominator. If the treatment changes who places orders, the denominator changes along with the numerator, and the ratio can fall while the business does better.

Example 5.7 (New small buyers)

Two groups of 100 users. In control, 10 users each place one \$50 order. In the treatment the same 10 buy, plus 10 new users who each place one \$20 order.

control revenue per user = $10 \times 50 / 100 = 5$	(total revenue over users)
treatment revenue per user = $(500 + 200) / 100 = 7$	(10 old and 10 new buyers)
control order value = $500 / 10 = 50$	(revenue over orders)
treatment order value = $700 / 20 = 35$	(revenue over orders)

The treatment earns 40% more per user (\$7 against \$5) while its average order value falls 30% (\$35 against \$50). Both statements are true. Only the per-user mean answers whether the treatment is better for the business, because its denominator, the users, is fixed by the randomization. A ratio of two outcomes is useful to explain a result and risky to decide on.

In ABMeter, a metric whose formula divides one measure by another is this ratio of means, so its value is a total over a total.

Exercises

1. Four users have revenue (40, 50, 0, 0) and orders (2, 1, 0, 0). Compute the mean of ratios and the ratio of means, and say how many users each uses.
2. A treatment makes every buyer place twice as many orders at the same order value and converts no new users. By what factor do revenue per user and average order value change?
3. Show that $M = R$ (the mean of ratios equals the ratio of means) when every user with $Y_i > 0$ has the same value of Y_i and every user with $Y_i = 0$ has $X_i = 0$.

The delta method

Chapter 5 made average order value a function of two means, $R = \bar{X}/\bar{Y}$ (the ratio of means). The central limit theorem gives the standard error of each mean, not of the ratio. The **delta method** fills the gap for any smooth function of means: near the true means the function is almost linear, and the standard error of a linear function follows from Chapter 1. Deng, Knoblich and Lu (2018) is a practical guide to the method for online experiments; its Section 3 treats exactly this case, where users are randomized and the metric is a ratio over things users do.

One variable

The idea

Zoom in far enough on a smooth curve and it looks like a straight line. The sample mean wobbles only a little around the true mean, so over that small range the curve f is almost its tangent line. A straight line with slope 3 turns a wobble of 0.1 into a wobble of 0.3, so the standard error of $f(\bar{X})$ is about the slope times the standard error of $\bar{X}$.

Why a derivative shows up in a standard error

Chapter 1 can handle the spread of $a\bar{X} + b$, a straight line through the mean, and nothing else. A curve such as $1/p$ is not a straight line, but the only part of it the sample mean ever visits is a short stretch around μ , and over a short stretch a smooth curve is one straight line, the tangent, with slope $f'(\mu)$. The delta method replaces the curve by that line and then uses Chapter 1. The derivative is simply the a in $a\bar{X} + b$: it says how much a small wobble in the input is stretched or squashed on the way to the output.

Theorem 6.1 (Delta method, one variable)

Let $\bar{X}$ be the mean of n i.i.d. observations with mean μ and variance σ^2 . By the central limit theorem, written with convergence in distribution ([Definition 2.7](#), read “converges in distribution to”),

$$\sqrt{n} \cdot (\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Let f be a function that is differentiable at μ with derivative $f'(\mu) \neq 0$. Then

$$\sqrt{n} \cdot (f(\bar{X}) - f(\mu)) \xrightarrow{d} N(0, f'(\mu)^2 \sigma^2)$$

so

$$\text{SE}(f(\bar{X})) \approx |f'(\mu)| \text{SE}(\bar{X})$$

Proof.

Here μ is the mean of one user’s value and σ^2 the variance of one user’s value. By Taylor’s theorem $f(\bar{X}) = f(\mu) + f'(\mu)(\bar{X} - \mu) + r$, where the remainder r is small compared with $|\bar{X} - \mu|$ as $\bar{X} \rightarrow \mu$. Multiplying by $\sqrt{n}$:

$$\sqrt{n} \cdot (f(\bar{X}) - f(\mu)) = f'(\mu) \cdot \sqrt{n} \cdot (\bar{X} - \mu) + \sqrt{n} \cdot r$$

As n grows, the first term on the right converges in distribution to $f'(\mu) \cdot W$, with $W \sim N(0, \sigma^2)$, by the central limit theorem. The second term goes to 0, because r is small compared with $|\bar{X} - \mu|$ while $\sqrt{n} \cdot |\bar{X} - \mu|$ stays of the order of σ . So

$$\sqrt{n} \cdot (f(\bar{X}) - f(\mu)) \xrightarrow{d} f'(\mu) \cdot W$$

and a constant times a normal variable is normal with the variance multiplied by the constant squared ([Proposition 1.5](#)):

$$f'(\mu) \cdot W \sim N(0, f'(\mu)^2 \sigma^2)$$

For the standard error, divide the limit variance by n and take the square root:

$$\text{SE}(f(\bar{X})) \approx \sqrt{f'(\mu)^2 \sigma^2 / n}$$

and the right side simplifies, with $\text{SE}(\bar{X}) = \sigma / \sqrt{n}$ ([Proposition 1.21](#)):

$$\begin{aligned} \sqrt{f'(\mu)^2 \sigma^2 / n} &= |f'(\mu)| \sigma / \sqrt{n} && \text{(take the square root of each factor)} \\ &= |f'(\mu)| \text{SE}(\bar{X}) && (\text{SE}(\bar{X}) = \sigma / \sqrt{n}) \end{aligned}$$

Example 6.2 (Visitors per signup)

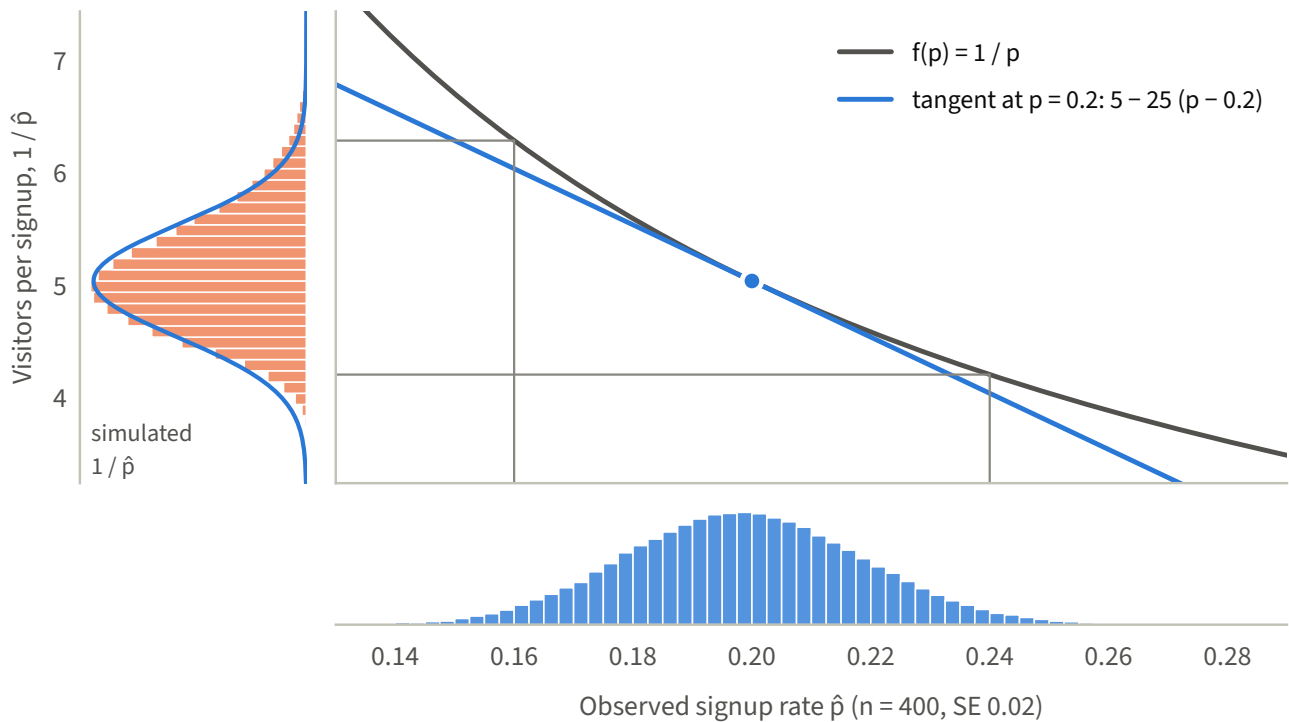
A page converts a share $p = 0.2$ of its visitors. Over $n = 400$ visitors the observed share $\hat{p}$ (the observed rate) is the sample mean of 0/1 values with variance $p(1 - p)$. The team reports visitors per signup, $f(p) = 1/p$.

$$\begin{aligned} \text{SE}(\hat{p}) &= \sqrt{p(1 - p)/n} && (\text{SE} = \sigma / \sqrt{n}, \text{ with } \sigma^2 = p(1 - p)) \\ &= \sqrt{0.2 \times 0.8 / 400} = 0.02 && \text{(substitute)} \\ f(p) &= 1/0.2 = 5 && \text{(visitors per signup)} \\ f'(p) &= -1/p^2 = -1/0.04 = -25 && \text{(derivative of } 1/p) \end{aligned}$$

The delta method gives $\text{SE}(1/\hat{p}) \approx |f'(\hat{p})| \text{SE}(\hat{p})$, and

$$|f'(\hat{p})| \text{SE}(\hat{p}) = 25 \times 0.02 = 0.5$$

In a simulation of 200,000 draws of $\hat{p}$, each the share of signups among 400 visitors converting at 0.2, the standard deviation of $1/\hat{p}$ is 0.515: 3% above the delta method's 0.5, because the curve $1/p$ bends upward and the tangent misses that. With 25 visitors the standard error of $\hat{p}$ is 0.08, the square root of $0.16 / 25$, and the same calculation gives $25 \times 0.08 = 2.0$, against a simulated 3.84: over the range where $\hat{p}$ now moves, roughly 0.04 to 0.36, the curve is far from straight.



The curve $1/p$ and its tangent at $p = 0.2$. The bottom histogram is the spread of $\hat{p}$ for 400 visitors; mapped through the curve it becomes the spread on the left (bars, simulation, 200,000 samples), which the delta method approximates by a normal curve with standard error 0.5 (line). The gray lines map $\hat{p} = 0.16$ and 0.24 through the curve.

The tangent is a good stand-in for the curve only where $\hat{p}$ actually moves. That range shrinks like $1/\sqrt{n}$, so the delta method is a large-sample method, like the central limit theorem it is built on.

Several means

The idea

With several averages the picture is the same, one dimension up: near the true means a smooth function is almost a flat plane, and its slopes in each direction form the gradient. The wobble of the result is a weighted sum of the wobbles of the inputs. Because the inputs come from the same users, their wobbles can move together, so the covariances enter the answer; they can make it larger or smaller.

Why the gradient, and why the cross term

A tangent plane in two variables has two slopes, one per direction, and those two numbers are the gradient. A small move (d_1, d_2) in the inputs changes the output by $g_1 d_1 + g_2 d_2$, a weighted sum, and the variance of a weighted sum was settled in Chapter 1: each weight squared times its variance, plus twice the product of the weights times the covariance ([Proposition 1.17](#)). The quadratic form $g^T \Sigma g$ is that formula written for k inputs at once. The cross term is not a refinement. When the two inputs move together and their weights have opposite signs, as in a ratio, it cancels most of the variance, and leaving it out can nearly double the standard error.

Theorem 6.3 (Delta method, several variables)

Let $m_1, \dots, m_k$ be k quantities measured on each user, with mean vector μ and covariance matrix Σ for one user ($\Sigma_{jl} = \text{Cov}(m_j, m_l)$). Let $\bar{m} = (\bar{m}_1, \dots, \bar{m}_k)$ be their sample means over n i.i.d. users, so that, by the multivariate central limit theorem (the same theorem for vectors, with $N(0, \Sigma)$ the multivariate normal distribution),

$$\sqrt{n} \cdot (\bar{m} - \mu) \xrightarrow{d} N(0, \Sigma)$$

If f is differentiable at μ with gradient $g = \nabla f(\mu) \neq 0$, the vector of its partial derivatives, then, with $g^\top$ the transpose of g ,

$$\sqrt{n} \cdot (f(\bar{m}) - f(\mu)) \xrightarrow{d} N(0, g^\top \Sigma g)$$

With the sample covariance matrix and the gradient at the sample means:

$$SE^2 = \frac{g^\top \Sigma g}{n}$$

Proof.

The proof is the one-variable proof with the derivative replaced by the gradient g (the gradient of f at the means): the linear term is $g_1(\bar{m}_1 - \mu_1) + \dots + g_k(\bar{m}_k - \mu_k)$. Its variance for one user is the quadratic form, with Σ the covariance matrix of the per-user measures:

$$\begin{aligned} \text{Var}(g_1 m_1 + \dots + g_k m_k) &= \sum_j \sum_l \text{Cov}(g_j m_j, g_l m_l) && \text{(Proposition 1.17, } k \text{ terms)} \\ &= \sum_j \sum_l g_j g_l \Sigma_{jl} && \text{(constants factor out)} \\ &= g^\top \Sigma g && \text{(definition of the matrix product)} \end{aligned}$$

Averaging over n users divides it by n ([Proposition 1.21](#)), which gives SE^2 . ■

Example 6.4 (Click-through rate in percent)

A page reports $f = 100 \cdot \bar{C}/\bar{V}$: clicks per view, in percent, from per-user means of clicks $\bar{C} = 1$ and views $\bar{V} = 4$ over $n = 50$ users. Suppose the per-user sample variances are $s_C^2 = 1$ and $s_V^2 = 4$ and the sample covariance is $s_{CV} = 1.5$: users who view more also click more.

$$\begin{aligned} f &= 100 \times 1/4 = 25 && \text{(value at the sample means)} \\ g &= (100/\bar{V}, -100 \bar{C}/\bar{V}^2) = (25, -6.25) && \text{(partial derivatives)} \\ g^\top \Sigma g &= 25^2 \times 1 + 2 \times 25 \times (-6.25) \times 1.5 + 6.25^2 \times 4 && \text{(quadratic form with } k = 2) \\ &= 625 - 468.75 + 156.25 = 312.5 && \text{(multiply and add)} \\ SE &= \sqrt{312.5/50} = 2.5 && \text{(Theorem 6.3)} \end{aligned}$$

The rate is 25% with a standard error of 2.5 points. Dropping the covariance term would give 3.95, the square root of $781.25 / 50$, far too wide: clicks and views rise together, so their errors partly cancel in the ratio.

Corollary 6.5 (Ratio of two means)

For $R = \bar{X}/\bar{Y}$ over n users, with sample variances s_X^2 and s_Y^2 and sample covariance s_{XY} , the gradient is $(1/\bar{Y}, -\bar{X}/\bar{Y}^2)$, and

$$SE_R^2 = \frac{s_X^2 - 2R s_{XY} + R^2 s_Y^2}{n \bar{Y}^2}$$

Proof.

Estimate Σ (the covariance matrix of the per-user measures) by s_X^2 , s_Y^2 and s_{XY} (the sample variances and the sample covariance), and evaluate the gradient at the sample means.

$$\begin{aligned} g &= (1/\bar{Y}, -\bar{X}/\bar{Y}^2) && \text{(partial derivatives of } x/y) \\ g^\top \Sigma g &= g_1^2 s_X^2 + 2g_1 g_2 s_{XY} + g_2^2 s_Y^2 && \text{(the quadratic form for } k=2) \\ &= \frac{s_X^2}{\bar{Y}^2} - 2 \frac{\bar{X}}{\bar{Y}^3} s_{XY} + \frac{\bar{X}^2}{\bar{Y}^4} s_Y^2 && \text{(substitute } g) \\ &= \frac{1}{\bar{Y}^2} \left(s_X^2 - 2 \frac{\bar{X}}{\bar{Y}} s_{XY} + \frac{\bar{X}^2}{\bar{Y}^2} s_Y^2 \right) && \text{(factor out } 1/\bar{Y}^2) \\ &= \frac{s_X^2 - 2R s_{XY} + R^2 s_Y^2}{\bar{Y}^2} && (R = \bar{X}/\bar{Y}) \\ SE_R^2 &= g^\top \Sigma g / n && \text{(Theorem 6.3)} \end{aligned}$$

Example 6.6 (Alice and Bob, by hand)

From Chapter 5: $X = (100, 90, 0, \dots, 0)$, $Y = (1, 9, 0, \dots, 0)$, $n = 10$, $\bar{X} = 19$, $\bar{Y} = 1$, $R = 19$ (the ratio of means).

- s_X^2 : deviations 81, 71 and eight times -19 ; squares $6561 + 5041 + 8 \times 361 = 14490$; divided by 9 gives 1610.
- s_Y^2 : deviations 0, 8 and eight times -1 ; squares $0 + 64 + 8 = 72$; divided by 9 gives 8.
- s_{XY} : products $81 \times 0 + 71 \times 8 + 8 \times 19 = 720$; divided by 9 gives 80.

$$\begin{aligned} s_X^2 - 2R s_{XY} + R^2 s_Y^2 &= 1610 - 2 \times 19 \times 80 + 19^2 \times 8 && \text{(substitute)} \\ &= 1610 - 3040 + 2888 && \text{(multiply)} \\ &= 1458 && \text{(add)} \\ SE_R &= \sqrt{\frac{1458}{10 \times 1^2}} && \text{(the corollary with } n=10, \bar{Y}=1) \\ &= \sqrt{145.8} = 12.07 && \text{(divide, take the root)} \end{aligned}$$

The two common mistakes give different answers. Dropping the covariance term leaves $(1610 + 2888) / 10 = 449.8$, whose root is 21.21, too large; treating the ten orders as independent observations gives 9.00 (Chapter 5), too small, because Bob's nine orders came from one randomized user.

Proposition 6.7 (Residual form)

The numerator of SE_R^2 is s_{X-RY}^2 , the sample variance of the residuals $X_i - R Y_i$, so, with s_{X-RY} its square root,

$$SE_R = \frac{s_{X-RY}}{\sqrt{n} \bar{Y}}$$

Proof.

Treat R (the ratio of means) as a constant. The rules of Chapter 1 hold for sample variances and covariances too, by the same algebra.

$$\begin{aligned}
 s_{X-RY}^2 &= s_X^2 + s_{-RY}^2 + 2 s_{X, -RY} && \text{(Proposition 1.17)} \\
 &= s_X^2 + R^2 s_Y^2 + 2 s_{X, -RY} && \text{(Proposition 1.5 with } a = -R) \\
 &= s_X^2 + R^2 s_Y^2 - 2R s_{XY} && \text{(a constant factors out of a covariance)}
 \end{aligned}$$

This is the numerator of SE_R^2 . ■

Example 6.8 (Alice and Bob's residuals)

With $R = 19$, the residuals are $100 - 19 \times 1 = 81$ for Alice, $90 - 19 \times 9 = -81$ for Bob, and 0 for the others. Their mean is 0, so their sample variance is $(81^2 + 81^2) / 9 = 1458$: the same numerator as in [Example 6.6](#). The residual form says what makes a ratio uncertain: users whose own ratio disagrees with the overall one. If every order were worth exactly R , every residual would be 0, and so would the standard error.

Computing the gradient

The idea

The gradient does not have to be worked out by hand. Any formula built from means, constants and the four arithmetic operations can carry, next to each intermediate value, its slope with respect to every input. Each operation updates the slope by the rule from calculus (sum, product or quotient rule), so at the end the formula's value and its gradient arrive together.

Any formula built from means, constants and $+ - \times \div$ is differentiable wherever no denominator is 0, and its gradient can be computed exactly in one pass over the formula's syntax tree. **Forward-mode automatic differentiation** carries a pair (value, gradient) through every node:

- a constant c gives $(c, 0)$, and mean j gives $(\bar{m}_j, e_j)$, with e_j the j -th unit vector;
- $(a, \nabla a) \pm (b, \nabla b) = (a \pm b, \nabla a \pm \nabla b)$;
- $(a, \nabla a) \times (b, \nabla b) = (ab, b \nabla a + a \nabla b)$, the product rule;
- $(a, \nabla a) \div (b, \nabla b) = (a/b, (b \nabla a - a \nabla b)/b^2)$, the quotient rule.

Example 6.9 (Forward mode on two formulas)

For $\bar{X}/\bar{Y}$ at $(19, 1)$, the leaves are $(19, [1, 0])$ and $(1, [0, 1])$, and the division returns 19 with gradient $([1, 0] \times 1 - 19 \times [0, 1]) / 1 = [1, -19]$, the gradient of the corollary. For $100 \cdot \bar{C}/\bar{V}$ at $(1, 4)$:

$$\begin{aligned}
 100 \cdot \bar{C} &\mapsto (100, [100, 0]) && \text{(constant times the leaf } (1, [1, 0])) \\
 \bar{V} &\mapsto (4, [0, 1]) && \text{(leaf)} \\
 \text{quotient} &\mapsto (100/4, ([100, 0] \times 4 - 100 \times [0, 1])/4^2) && \text{(quotient rule)} \\
 &= (25, [25, -6.25]) && \text{(evaluate)}
 \end{aligned}$$

the value and the gradient used in the click-through example.

Why the covariance matters

The idea

Revenue and orders come from the same users and rise together. When a sample happens to contain more big spenders, both totals go up, and their ratio moves much less than either total. Dropping the covariance ignores this cancellation and overstates the uncertainty.

Example 6.10 (Coverage with and without the covariance)

A simulation draws 4,000 samples at each size from a population with a Poisson number of orders per user (mean 0.4) and lognormal order values with log-mean 3.5 and log-standard deviation 0.8. The true average order value is the lognormal mean, $e^{3.5+0.8^2/2} = \$45.60$, and the simulation counts how often each 95% interval contains it:

users	delta method: coverage	mean width	covariance dropped: coverage	mean width
100	90.0%	25.17	99.8%	47.70
1,000	93.8%	8.37	99.9%	15.18
10,000	94.8%	2.68	100.0%	4.81

Naive: covariance dropped

The interval covers the truth almost always (99.9% at 1,000 users) because it is about 80% wider than it should be (mean width 15.18 against 8.37). A test built on it misses effects it could detect.

Correct: the delta method with the covariance

Coverage approaches 95% as n grows: 90.0% at 100 users, 93.8% at 1,000 and 94.8% at 10,000. The shortfall at small n is the large-sample nature of the method.

Checking with the bootstrap

The idea

To see how much a statistic would wobble if you re-ran the experiment, you can fake the re-runs: draw a new sample of the same size from your own users, with replacement, recompute the statistic, and repeat many times. The spread of the recomputed values approximates the standard error, with no formula needed.

Definition 6.11 (Bootstrap standard error)

Draw n users with replacement from the sample, recompute the statistic, and repeat B times. The standard deviation of the B values is the **bootstrap standard error** (Efron and Tibshirani 1993).

Example 6.12 (Every resample of three users)

Three users spent 1, 2 and 6. There are $3^3 = 27$ equally likely resamples of size 3, and their means have the variance of a mean of 3 draws from the values 1, 2, 6:

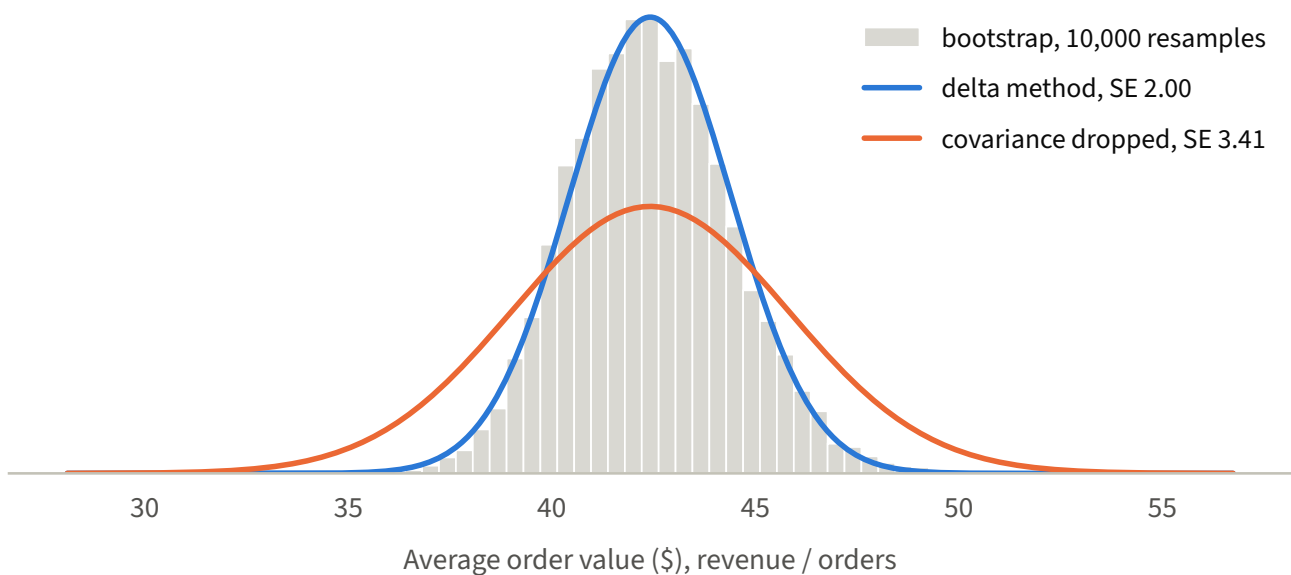
$$\begin{aligned} \text{variance of the values} &= (4 + 1 + 9)/3 = 14/3 && \text{(squared deviations from 3, over } n) \\ \text{variance of a resample mean} &= (14/3)/3 = 14/9 && \text{(Proposition 1.21)} \\ \text{bootstrap SE} &= \sqrt{14/9} = 1.25 && \text{(square root)} \end{aligned}$$

The formula $s/\sqrt{n}$ gives $\sqrt{7} / \sqrt{3} = 1.53$, since $s^2 = 14/2 = 7$ (the sample variance). The bootstrap divides by n instead of $n - 1$, which matters only in tiny samples.

The bootstrap needs no formula, so it is an independent check on the delta method, at the price of B recomputations.

Example 6.13 (Bootstrap against the delta method)

For one simulated sample of 1,000 users from the population of the coverage simulation (Poisson(0.4) orders, lognormal order values) and 10,000 resamples, the delta-method standard error of average order value is 1.995 and the bootstrap gives 1.987; for the product of revenue per user and orders per user, 0.842 and 0.841. For a sample of 5,000 users the two agree within 1%. At Alice and Bob's 10 users the delta method gives 12.07 and the bootstrap 37.32: the delta method understates the spread by two thirds. In about one resample in nine neither Alice nor Bob is drawn (all ten draws miss both with probability $0.8^{10} = 0.107$), so there are no orders and no ratio at all.



Average order value at $n = 1,000$: the bootstrap distribution (gray, 10,000 resamples) against the delta method's normal curve (blue) and the curve with the covariance dropped (orange).

Bias of a ratio

The idea

Dividing by a noisy number is lopsided. If the denominator is sometimes 0.5 and sometimes 1.5 around a true value of 1, one over it is sometimes 2 and sometimes 0.67: on average 1.33, not 1. So a ratio of two unbiased means is biased upward by an amount that grows with the noise of the denominator, and shrinks like $1/n$.

Why a tangent cannot see a bias

A tangent line is symmetric: it stretches a wobble above the mean and a wobble below it by the same factor, so a linear approximation of an unbiased mean is itself unbiased. The curve $1/y$ is not symmetric; it rises more steeply on the left than it falls on the right. To see the bias, the approximation needs the next term of Taylor's formula, the one with the second derivative, which measures that bending. The second derivative is multiplied by the squared wobble, whose average is the variance, and the variance of a mean is of order $1/n$: that is why the bias shrinks like $1/n$ while the standard error shrinks like $1/\sqrt{n}$.

Proposition 6.14 (Second-order bias of a ratio)

Let the pairs (X_i, Y_i) be i.i.d. over n users with means μ_X and $\mu_Y \neq 0$, per-user variance $\sigma_Y^2 = \text{Var}(Y_i)$ and covariance $\sigma_{XY} = \text{Cov}(X_i, Y_i)$. The ratio of means $R = \bar{X}/\bar{Y}$ estimates μ_X/μ_Y with bias

$$\mathbb{E}[R] - \frac{\mu_X}{\mu_Y} \approx \frac{\mu_X \sigma_Y^2 - \mu_Y \sigma_{XY}}{n \mu_Y^3}$$

Proof.

Write $d_X = \bar{X} - \mu_X$ and $d_Y = \bar{Y} - \mu_Y$ for the errors of the two sample means. The derivatives of x/y at (μ_X, μ_Y) are $1/\mu_Y$ and $-\mu_X/\mu_Y^2$ (first order), and 0 , $-1/\mu_Y^2$ and $2\mu_X/\mu_Y^3$ (second order: in x twice, mixed, in y twice). Taylor's formula to second order halves each second derivative and counts the mixed one twice:

$$\frac{\bar{X}}{\bar{Y}} - \frac{\mu_X}{\mu_Y} \approx \frac{d_X}{\mu_Y} - \frac{\mu_X d_Y}{\mu_Y^2} - \frac{d_X d_Y}{\mu_Y^2} + \frac{\mu_X d_Y^2}{\mu_Y^3}$$

Take expectations. The first two terms vanish, since $\mathbb{E}[d_X] = \mathbb{E}[d_Y] = 0$:

$$\mathbb{E}\left[\frac{\bar{X}}{\bar{Y}}\right] - \frac{\mu_X}{\mu_Y} \approx -\frac{\mathbb{E}[d_X d_Y]}{\mu_Y^2} + \frac{\mu_X \mathbb{E}[d_Y^2]}{\mu_Y^3}$$

The right side equals

$$\begin{aligned} -\frac{\mathbb{E}[d_X d_Y]}{\mu_Y^2} + \frac{\mu_X \mathbb{E}[d_Y^2]}{\mu_Y^3} &= -\frac{\text{Cov}(\bar{X}, \bar{Y})}{\mu_Y^2} + \frac{\mu_X \text{Var}(\bar{Y})}{\mu_Y^3} && \text{(definitions)} \\ &= -\frac{\sigma_{XY}}{n \mu_Y^2} + \frac{\mu_X \sigma_Y^2}{n \mu_Y^3} && \text{(Proposition 1.21; the same for Cov)} \\ &= \frac{\mu_X \sigma_Y^2 - \mu_Y \sigma_{XY}}{n \mu_Y^3} && \text{(common denominator)} \end{aligned}$$

■

Example 6.15 (A noisy denominator)

Let the numerator mean be exactly $\bar{X} = 1$ and the denominator mean $\bar{Y}$ be 0.5 or 1.5 with probability $\frac{1}{2}$ each, so $\mu_Y = 1$ and $\text{Var}(\bar{Y}) = 0.25$.

$$\begin{aligned}\mathbb{E}[1/\bar{Y}] &= (1/0.5 + 1/1.5)/2 && \text{(the two equally likely values)} \\ &= (2 + 0.667)/2 = 1.333 && \text{(divide and average)}\end{aligned}$$

The second-order formula, with no covariance, predicts a bias of $\mu_X \text{Var}(\bar{Y})/\mu_Y^3 = 0.25$, so 1.25 against the exact 1.333: the right direction and roughly the right size even for this very noisy denominator.

For a realistic case, take a population where users with more orders place smaller orders, so that revenue is not proportional to orders: Poisson(0.4) orders per user, and an order value whose logarithm is normal with mean $4.0 - 0.4 \times \text{orders}$ and standard deviation 0.6. Its exact moments are $\mu_Y = 0.4$, $\sigma_Y^2 = 0.4$ (a Poisson variance equals its mean), $\mu_X/\mu_Y = 38.403$ and $\sigma_{XY} = 13.335$, so at $n = 20$ the formula gives

$$\frac{\mu_X \sigma_Y^2 - \mu_Y \sigma_{XY}}{n \mu_Y^3} = \frac{38.403 \times 0.4 - 13.335}{20 \times 0.16} = \frac{2.026}{3.2} = 0.633$$

(after dividing numerator and denominator by μ_Y). One million simulated samples per size give an average bias of +0.652 (± 0.012) at 20 users, +0.258 at 50 and +0.069 at 200, against +0.633, +0.253 and +0.063 from the formula; the formula is second order, so the gap shrinks faster than the bias itself.

The bias shrinks like $1/n$ and the standard error like $1/\sqrt{n}$, so for large samples the bias is negligible next to the interval width.

In ABMeter, the standard error of every metric whose formula divides or multiplies measures comes from this method, with the gradient computed by forward-mode differentiation.

Exercises

1. A metric is $2\bar{X}$. What is its gradient, and how does its standard error compare with that of $\bar{X}$?
2. For the product $f = \bar{C} \cdot \bar{V}$ of two per-user means at $\bar{C} = 4$ and $\bar{V} = 3$, give the value and the gradient, and write $\mathbf{g}^T \Sigma \mathbf{g}$ in terms of s_C^2 , s_V^2 and s_{CV} (the sample variances and the sample covariance).
3. Alice and Bob's data give $s_X^2 = 1610$, $s_Y^2 = 8$, $s_{XY} = 80$ and $R = 19$. If the covariance were -80 instead of 80 , what would SE_R be with $n = 10$ and $\bar{Y} = 1$?

The relative lift

Results are usually read as percentages: “+25% conversions” means more than “+0.0277”. The relative lift divides by the control mean, and the control mean is an estimate with its own error. This chapter derives the delta-method interval for the relative lift, compares it with the naive interval and with Fieller’s exact construction, and measures the small-sample bias.

The delta-method interval

The idea

A relative lift is “treatment divided by control, minus one”. Both the numerator and the denominator are noisy averages, so the percentage has two sources of uncertainty. Dividing an interval for the absolute difference by the control mean pretends the denominator is exact, and gets the width wrong whenever the lift is not zero.

Definition 7.1 (Relative lift)

For control and treatment means $\mu_C \neq 0$ and μ_T , the **relative lift** is $L = \mu_T/\mu_C - 1$. It is estimated by $\hat{L} = \bar{X}_T/\bar{X}_C - 1$, with $\bar{X}_T$ and $\bar{X}_C$ the sample means of the two groups.

Example 7.2 (A pricing-page lift)

On a pricing-page test, control users spend \$10 on average and treatment users \$12, each group with 100 users and sample variance 100. Then $\hat{L} = 12/10 - 1 = 0.2$, a 20% lift, and each group’s standard error is 1, the square root of $100 / 100$.

Proposition 7.3 (Standard error of the relative lift)

For independent groups with sample means $\bar{X}_T$ and $\bar{X}_C$ and standard errors SE_T and SE_C of those means,

$$SE_{\text{rel}}^2 = \frac{SE_T^2}{\bar{X}_C^2} + \frac{\bar{X}_T^2 SE_C^2}{\bar{X}_C^4}$$

and the 95% interval is $\hat{L} \pm 1.96 SE_{\text{rel}}$.

Proof.

Apply [Theorem 6.3](#) to $f(t, c) = t/c - 1$ at $t = \bar{X}_T$ and $c = \bar{X}_C$ (the treatment and control sample means). The two means come from different users, so the covariance matrix of $(\bar{X}_T, \bar{X}_C)$, which takes the place of Σ/n , is diagonal with entries SE_T^2 and SE_C^2 (the standard errors of the treatment and control means, squared).

$$\begin{aligned} g &= (1/c, -t/c^2) && \text{(partial derivatives of } t/c - 1) \\ SE_{\text{rel}}^2 &= g_1^2 SE_T^2 + g_2^2 SE_C^2 && \text{(quadratic form with a diagonal matrix)} \\ &= \frac{SE_T^2}{\bar{X}_C^2} + \frac{\bar{X}_T^2 SE_C^2}{\bar{X}_C^4} && \text{(substitute } g) \end{aligned}$$

■

The naive alternative takes the interval of the absolute difference and divides both ends by the control mean $\bar{X}_C$. Its implied standard error is

$$\begin{aligned} \text{SE}_{\text{naive}}^2 &= (\text{SE}_T^2 + \text{SE}_C^2) / \bar{X}_C^2 && \text{(standard error of the difference, divided by } \bar{X}_C) \\ &= \frac{\text{SE}_T^2}{\bar{X}_C^2} + \frac{\text{SE}_C^2}{\bar{X}_C^2} && \text{(split the fraction)} \end{aligned}$$

The delta method's control term is $\bar{X}_T^2 \text{SE}_C^2 / \bar{X}_C^4 = (\bar{X}_T / \bar{X}_C)^2 \cdot \text{SE}_C^2 / \bar{X}_C^2$, so the two differ only in the control term, which the delta method multiplies by $(\bar{X}_T / \bar{X}_C)^2$. When the treatment is above control the naive interval is too narrow; below control it is too wide; at zero lift they agree.

Example 7.4 (The two intervals for the pricing page)

For [Example 7.2](#): $\bar{X}_T = 12$, $\bar{X}_C = 10$, $\text{SE}_T = \text{SE}_C = 1$ (the standard errors of the treatment and control means).

Naive: absolute interval divided by control

The difference has standard error $\sqrt{2}$, the square root of $1 + 1$, so the absolute interval is $2 \pm 1.96 \times \sqrt{2} = [-0.7719, 4.7719]$; divided by 10 it is $[-0.0772, 0.4772]$.

Correct: the delta method

$$\begin{aligned} \text{SE}_{\text{rel}} &= \sqrt{\frac{1}{10^2} + \frac{12^2 \times 1}{10^4}} && \text{(the proposition with } \text{SE}_T = \text{SE}_C = 1) \\ &= \sqrt{0.01 + 0.0144} && \text{(divide)} \\ &= \sqrt{0.0244} = 0.1562 && \text{(add, take the root)} \end{aligned}$$

so the interval is $0.2 \pm 1.96 \times 0.1562 = [-0.1062, 0.5062]$, about 11% wider than the naive one (0.1562 against $\sqrt{2} / 10 = 0.1414$).

Fieller's interval

The idea

The delta-method interval is symmetric around $\hat{L}$, but a ratio's distribution is lopsided: a control mean that happens to be low inflates the ratio more than a high one deflates it. Fieller's idea avoids the approximation. For a candidate true ratio θ , the combination $\bar{X}_T - \theta \bar{X}_C$ should look like pure noise around 0. Keep every θ for which it does.

The step Fieller removed

The delta method replaces a ratio by its tangent and then borrows the normal distribution for the result. Fieller (1954) asked for an interval with no tangent in it. His observation: a ratio of two normal means is awkward, but $\bar{X}_T - \theta \bar{X}_C$ for a fixed candidate θ is a plain linear combination of them, whose distribution is exact. Test every candidate with that combination and keep the ones that pass. The price is a quadratic equation instead of a formula; the gain is an interval that follows the lopsided shape of a ratio's error instead of forcing symmetry on it.

Proposition 7.5 (Fieller's interval)

Let $\theta = \mu_T/\mu_C$ (the true ratio of the treatment and control means), and suppose $\bar{X}_T$ and $\bar{X}_C$ are independent and normal with known standard errors SE_T and SE_C . Then $D = \bar{X}_T - \theta\bar{X}_C$ is normal, with

$$\begin{aligned}\mathbb{E}[D] &= \mu_T - \theta\mu_C && \text{(linearity)} \\ &= 0 && (\theta = \mu_T/\mu_C) \\ \text{Var}(D) &= SE_T^2 + \text{Var}(\theta\bar{X}_C) && \text{(independent: variances add)} \\ &= SE_T^2 + \theta^2 SE_C^2 && \text{(Proposition 1.5)}\end{aligned}$$

So D divided by its standard deviation is standard normal, and the set of θ with $D^2 \leq 1.96^2 \text{Var}(D)$, that is

$$(\bar{X}_T - \theta\bar{X}_C)^2 \leq 1.96^2(SE_T^2 + \theta^2 SE_C^2)$$

contains the true ratio with probability exactly 0.95. It is a quadratic inequality in θ , and the set is a finite interval when $\bar{X}_C^2 > 1.96^2 SE_C^2$, that is, when the control mean is itself significantly different from 0. Then the coefficient of θ^2 , $\bar{X}_C^2 - 1.96^2 SE_C^2$, is positive, and the set lies between the two roots.

Example 7.6 (Fieller for the pricing page)

For [Example 7.2](#) ($\bar{X}_T = 12$, $\bar{X}_C = 10$, $SE_T = SE_C = 1$):

$$\begin{aligned}(12 - 10\theta)^2 &\leq 1.96^2(1 + \theta^2) && \text{(Fieller's inequality)} \\ 144 - 240\theta + 100\theta^2 &\leq 3.8416 + 3.8416\theta^2 && \text{(expand both sides)} \\ 96.1584\theta^2 - 240\theta + 140.1584 &\leq 0 && \text{(move everything to the left)}\end{aligned}$$

The quadratic formula gives the roots:

$$\begin{aligned}\theta &= \frac{240 \pm \sqrt{240^2 - 4 \times 96.1584 \times 140.1584}}{2 \times 96.1584} && \text{(quadratic formula)} \\ &= \frac{240 \pm 60.75}{192.32} && \text{(evaluate the root)} \\ &= 0.9321 \text{ or } 1.5638 && \text{(divide)}\end{aligned}$$

The coefficient of θ^2 is positive, so the set is the interval between the roots, and subtracting 1 gives the relative-lift interval $[-0.0679, 0.5638]$: shifted to the right of the delta interval $[-0.1062, 0.5062]$, as the lopsidedness predicts.

Deng, Knoblich and Lu (2018, Section 2 and Table 1) compare the two by simulation: from about 200 units per group both reach their nominal coverage, and Fieller is ahead only in small samples.

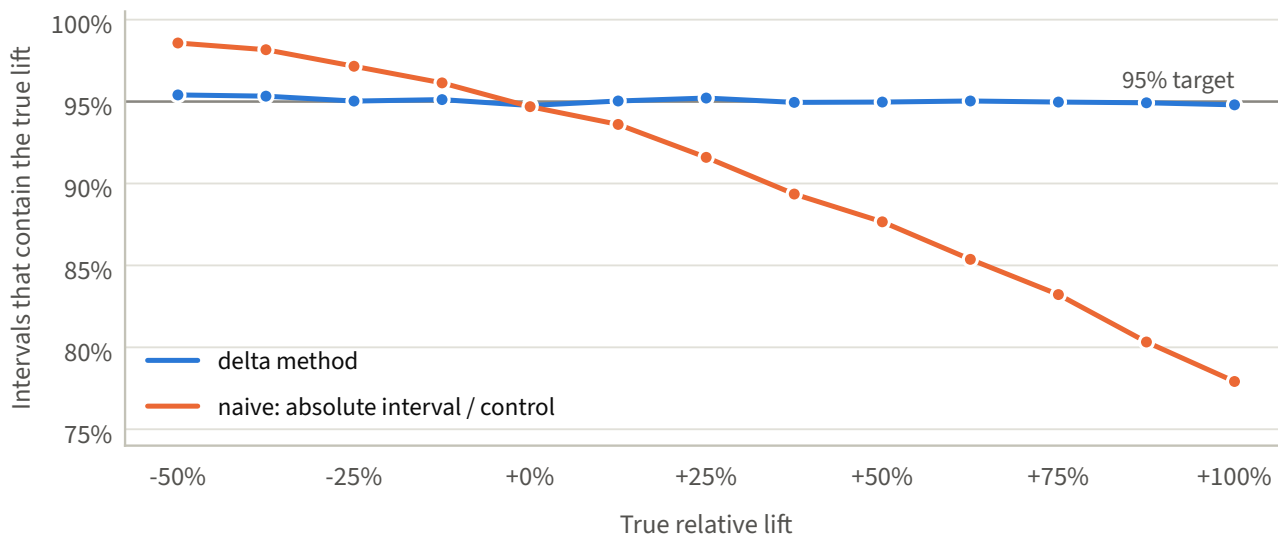
Coverage

Example 7.7 (Coverage of the three intervals)

A simulation runs 20,000 experiments for each true lift, with normal data, standard deviation 10 and a control mean of 10, and counts how often each 95% interval contains the true lift:

users per group	true lift	naive	delta	Fieller
100	-50%	98.6%	95.4%	94.9%
100	0%	94.8%	94.9%	94.8%
100	+50%	87.4%	94.8%	94.5%
100	+100%	78.5%	95.0%	95.1%
10	+100%	76.8%	91.3%	93.3%

With 100 users per group the delta and Fieller intervals both hold 95%. The naive interval covers a true doubling only 78.5% of the time: it is most overconfident about exactly the large gains a team hopes for. With 10 users per group Fieller's interval (93.3%) is closer to 95% than the delta interval (91.3%).



Simulation, 20,000 experiments per lift: coverage of the relative-lift interval against the true lift, 100 users per group. The delta-method interval holds 95% at every lift; the naive interval is right only at zero lift, too wide for losses and too narrow for gains.

Bias of the relative lift

The idea

The relative lift is a ratio, so it inherits the upward bias of Chapter 6: a noisy control mean in the denominator makes the estimated lift too high on average. The bias is small, of order $1/n$, and a second-order correction removes most of it.

Proposition 7.8 (Bias of a ratio of independent means)

Let $\bar{X}_T$ and $\bar{X}_C$ be independent, with true means μ_T and μ_C and SE_C the standard error of the control mean. Then

$$\mathbb{E} \left[\frac{\bar{X}_T}{\bar{X}_C} \right] - \frac{\mu_T}{\mu_C} \approx \frac{\mu_T}{\mu_C} \cdot \frac{SE_C^2}{\mu_C^2}$$

so the estimated lift is too high on average, and the corrected estimate subtracts this term.

Proof.

[Proposition 6.14](#) applied with X the treatment and Y the control gives, since the two means come from different users and their covariance term is 0,

$$\mathbb{E} \left[\frac{\bar{X}_T}{\bar{X}_C} \right] - \frac{\mu_T}{\mu_C} \approx \frac{\mu_T \sigma_C^2}{n \mu_C^3}$$

with σ_C^2 the variance of one control user's value. The right side equals

$$\begin{aligned} \frac{\mu_T \sigma_C^2}{n \mu_C^3} &= \frac{\mu_T}{\mu_C} \cdot \frac{\sigma_C^2/n}{\mu_C^2} && \text{(regroup)} \\ &= \frac{\mu_T}{\mu_C} \cdot \frac{SE_C^2}{\mu_C^2} && (SE_C^2 = \sigma_C^2/n) \end{aligned}$$

■

Example 7.9 (Correcting the pricing-page lift)

For [Example 7.2](#), with ratio 1.2, $SE_C^2 = 1$ (the squared standard error of the control mean) and $\bar{X}_C^2 = 100$:

$$\begin{aligned} \text{bias term} &= 1.2 \times 1/100 = 0.012 && \text{(the proposition, with estimates in place of true values)} \\ \text{corrected lift} &= 0.2 - 0.012 = 0.188 && \text{(subtract the bias)} \end{aligned}$$

A simulation of 2,000,000 experiments with 50 users per group (means 10 and 12, standard deviation 10, so $SE_C^2 = 2$) shows the bias directly: the average estimated ratio is 1.2254, high by 0.0254, while the second-order term is $1.2 \times 2 / 100 = 0.0240$ (ratio times SE_C^2/μ_C^2). Subtracting the term brings the average to 1.1980; adding it would give 1.2528.

Remark 7.10 (A sign in Deng, Knoblich and Lu (2018))

Their Algorithm 1 computes this same term and writes the corrected estimate as the estimate plus the term. The expansion above, and the simulation, show that the plain estimate is already high by the term, so the correction should subtract it. The difference is of order $1/n$ and does not matter for large groups.

In ABMeter, the percentage interval shown for each treatment group is the delta-method interval of this chapter, without the bias correction.

Exercises

- Control has mean 0.10 and treatment 0.08, each with standard error 0.01. Compute the relative lift, the delta-method 95% interval and the naive interval. Which is wider, and why?
- Show that the delta-method and naive intervals coincide when $\bar{X}_T = \bar{X}_C$ (the treatment and control sample means are equal).

3. Control has mean 10 with standard error 6. Is Fieller's interval for the ratio a finite interval?

Repeated looks, many comparisons and sample size

A level-0.05 test makes a false positive 5% of the time if it is computed once. Real experiments are watched daily, often have several treatments, and need a planned length. Each changes what “ $p < 0.05$ ” is worth. This chapter quantifies the first two and shows how to choose the sample size in advance.

Peeking

The idea

Every time you look at a running test, noise gets another chance to cross the significance line. If you stop the first time it does, you are no longer running one test at level 0.05; you are running many, and keeping the luckiest. Look often enough and a test where nothing changed will cross the line sooner or later.

Definition 8.1 (Fixed-horizon test)

A **fixed-horizon** test decides its sample size or duration before the data arrive and computes the test once, at the end. [Proposition 3.9](#) holds only under this rule.

Example 8.2 (Two looks instead of one)

A test of a signup form is read after one week and after two. If the two reads were independent, each with a 5% false-positive rate under H_0 (the null hypothesis), the chance that at least one is significant would be $1 - 0.95^2 = 0.0975$. The reads share the first week’s data, so they are positively correlated and the real rate is a little lower: 8.4% in the simulation below. Either way, two looks nearly double the false positives of one.

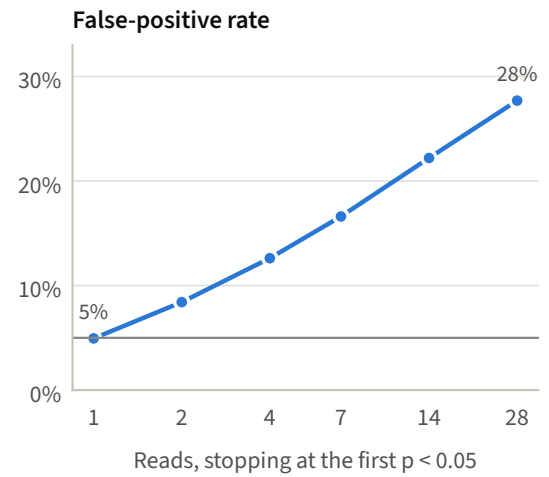
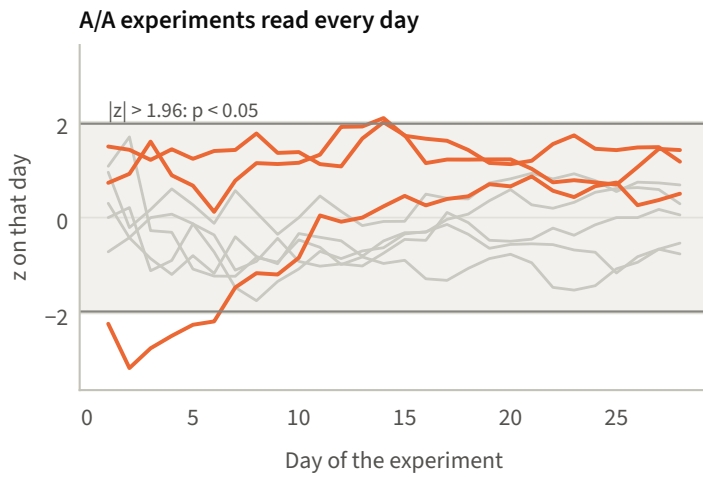
If the same experiment is tested every day and stopped at the first $p < 0.05$, a false positive occurs whenever the path of z (the test statistic) crosses ± 1.96 on any day. Under H_0 , z computed on growing data behaves like a random walk rescaled by its standard deviation, and such a path, given unlimited looks, crosses any fixed boundary with probability 1. Johari, Koomen, Pekelis and Walsh (2017) document the problem in A/B testing practice.

Example 8.3 (Daily looks at A/A tests)

A simulation runs 20,000 A/A experiments over 28 days, with 250 new users per group per day and a 10% conversion rate in both groups. Each is read 1 to 28 times at evenly spaced days and counted as a false positive if any read has $p < 0.05$:

reads	1	2	4	7	14	28
false-positive rate	4.9%	8.4%	12.6%	16.6%	22.2%	27.7%

Reading daily for four weeks turns 5% into 28%.



Simulation, 20,000 A/A experiments. Left: the daily z of several of them. Orange paths cross ± 1.96 on some day and end inside the band, so stopping at the first crossing declares a difference that does not exist. Right: the false-positive rate grows with the number of reads.

Two remedies exist: fix the horizon and read once, or use a sequential method built for continuous monitoring, such as the always-valid p -values of Johari, Pekelis and Walsh (2015), which stay valid at any stopping time at the cost of some power.

Several comparisons

The idea

Each comparison at level 0.05 has a 1-in-20 chance of a false positive. Run twenty comparisons where nothing works, and you should expect about one “win”. The fix is to make each comparison stricter, so that the chance of any false positive across the whole family stays at 5%.

Proposition 8.4 (Bonferroni bound)

If k tests are each run at level α , the probability that at least one makes a type I error is at most $k\alpha$. Running each at level α/k keeps that probability at most α .

Proof.

Let A_i be the event that test i makes a type I error, so $P(A_i) \leq \alpha$ (the test level).

$$\begin{aligned} P(A_1 \cup \dots \cup A_k) &\leq P(A_1) + \dots + P(A_k) && \text{(a union is at most the sum of its parts)} \\ &\leq k\alpha && (P(A_i) \leq \alpha) \end{aligned}$$

At level α/k each $P(A_i) \leq \alpha/k$, and the same two steps give at most $k \cdot \alpha/k = \alpha$. ■

Example 8.5 (Three button colors)

A test compares three new checkout button colors with the current one, at a family-wise level of 0.05. Bonferroni runs each comparison at $0.05 / 3 = 0.0167$. If all three colors do nothing:

$$\begin{aligned}\text{Bonferroni bound} &= 3 \times 0.05 = 0.15 && \text{(each at 0.05, no correction)} \\ \text{independent tests} &= 1 - 0.95^3 = 0.143 && \text{(all three stay quiet with probability } 0.95^3\text{)}\end{aligned}$$

The comparisons share one control group, so they are not independent: when the control's rate happens to come out low, all three treatments look good at once. In 50,000 simulated experiments with 5,000 users per group converting at 10% in all four groups, at least one of the three comparisons is significant 12.5% of the time, below the 14.3% for independent tests because the comparisons are positively correlated.

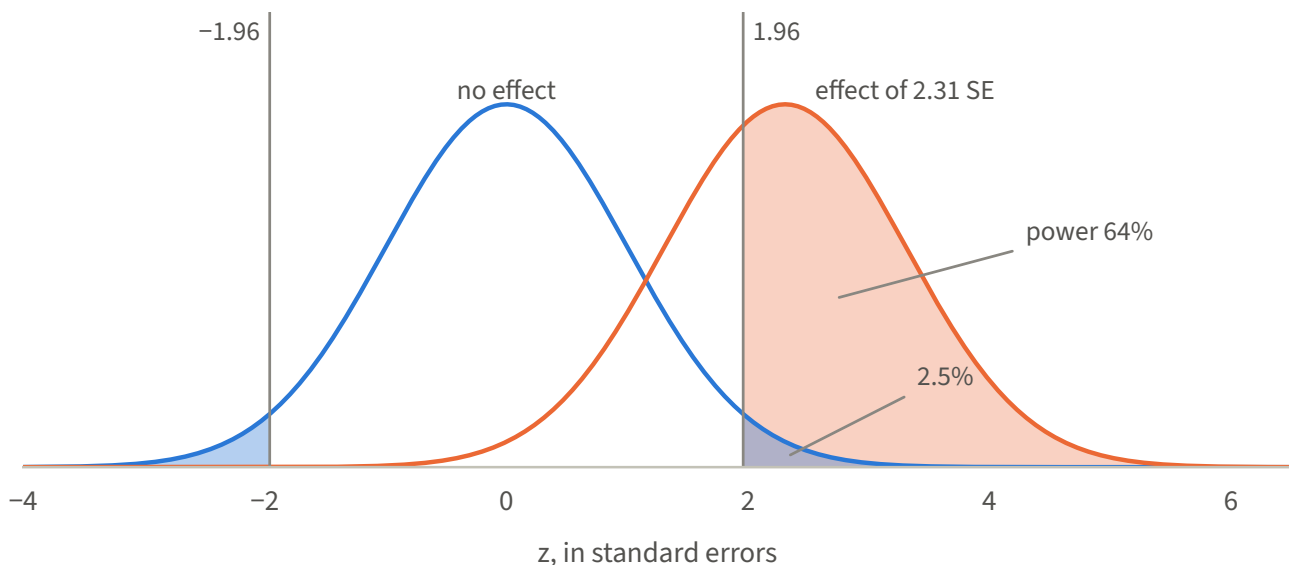
Power and sample size

The idea

Power is the chance of seeing an effect that is really there. What matters is the size of the effect measured in standard errors: an effect two standard errors wide is hard to miss, one standard error wide is a coin flip. More users shrink the standard error, so the sample size is chosen to make the effect you care about wide enough, in those units, to be found most of the time.

Two bells

Under H_0 the z-statistic is a bell centred at 0, and the test rejects outside ± 1.96 . If the true effect is δ , the statistic is the same bell slid to the right by δ/SE . Power is the part of the slid bell that lies beyond 1.96, shown in the figure. Everything about sample size follows from the picture. The slide grows like $\sqrt{n}$, so more users move the second bell to the right and more of it crosses the line; a larger effect moves it further for the same n ; a stricter level moves the line to the right and cuts the area.



The checkout test of the next example, 10,000 users per group. Blue: the distribution of z if the new checkout does nothing, with the 5% rejection region beyond ± 1.96 shaded. Orange: the same bell slid right by $\delta/\text{SE} = 2.31$ when the new checkout converts 11% against 10%. The shaded orange area, 64%, is the power: the share of such experiments that reject.

Proposition 8.6 (Power of the two-sample z-test)

If the true difference is δ and the standard error of the estimated difference is SE, the power $1 - \beta$ of the two-sided level- α test, with β the type II error rate and $z_{1-\alpha/2}$ the standard normal quantile (Definition 2.5), is approximately

$$1 - \beta \approx \Phi\left(\frac{|\delta|}{\text{SE}} - z_{1-\alpha/2}\right)$$

Proof.

Take $\delta > 0$; the case $\delta < 0$ is symmetric. The estimated difference is approximately $N(\delta, \text{SE}^2)$, so z (the test statistic) is approximately $N(\delta/\text{SE}, 1)$, and $z - \delta/\text{SE}$ is approximately standard normal. The test rejects when $|z| > z_{1-\alpha/2}$, and the lower tail, $z < -z_{1-\alpha/2}$, has negligible probability when $\delta > 0$. So the power is approximately

$$P(z > z_{1-\alpha/2})$$

and this probability equals

$$\begin{aligned} P(z > z_{1-\alpha/2}) &= P(z - \delta/\text{SE} > z_{1-\alpha/2} - \delta/\text{SE}) && \text{(subtract } \delta/\text{SE} \text{ on both sides)} \\ &= 1 - \Phi(z_{1-\alpha/2} - \delta/\text{SE}) && (z - \delta/\text{SE} \text{ is standard normal)} \\ &= \Phi(\delta/\text{SE} - z_{1-\alpha/2}) && \text{(symmetry: } 1 - \Phi(a) = \Phi(-a)) \end{aligned}$$

Example 8.7 (Power of a checkout test)

The current checkout converts 10% and the new one 11%, so $\delta = 0.01$ (the true difference), with per-user variances $0.1 \times 0.9 = 0.09$ and $0.11 \times 0.89 = 0.0979$. With 10,000 users per group and $\Phi(0.347) = 0.636$:

$$\begin{aligned} \text{SE} &= \sqrt{(0.09 + 0.0979)/10,000} = 0.00433 && \text{(Proposition 1.21 in each group)} \\ \delta/\text{SE} &= 0.01/0.00433 = 2.307 && \text{(the effect in standard errors)} \\ \Phi(\delta/\text{SE} - 1.96) &= \Phi(2.307 - 1.96) = \Phi(0.347) = 0.636 && \text{(power, with } z_{0.975} = 1.96) \end{aligned}$$

A real 1-point lift would be detected only about 64% of the time.

Corollary 8.8 (Sample size)

With n users per group and per-user variances σ_C^2 and σ_T^2 , the number of users per group needed for power $1 - \beta$ is

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (\sigma_C^2 + \sigma_T^2)}{\delta^2}$$

Proof.

By Proposition 1.21 in each group, $\text{SE}^2 = (\sigma_C^2 + \sigma_T^2)/n$. Setting the power to $1 - \beta$ in the proposition, and using $\Phi(z_{1-\beta}) = 1 - \beta$:

$$\begin{aligned} |\delta|/\text{SE} - z_{1-\alpha/2} &= z_{1-\beta} && (\Phi \text{ is increasing, so the arguments are equal)} \\ \text{SE} &= |\delta|/(z_{1-\alpha/2} + z_{1-\beta}) && \text{(solve for SE)} \\ (\sigma_C^2 + \sigma_T^2)/n &= \delta^2/(z_{1-\alpha/2} + z_{1-\beta})^2 && \text{(square both sides; } \text{SE}^2 = (\sigma_C^2 + \sigma_T^2)/n) \\ n &= (z_{1-\alpha/2} + z_{1-\beta})^2 (\sigma_C^2 + \sigma_T^2)/\delta^2 && \text{(solve for } n) \end{aligned}$$

For $\alpha = 0.05$ and 80% power, $z_{0.975} = 1.96$ and $z_{0.8} = 0.8416$ (standard normal quantiles), so $(1.96 + 0.8416)^2 = 7.849$. With equal variances σ^2 in both groups the numerator is $7.849 \times 2\sigma^2 \approx 15.7\sigma^2$, so n is about $16\sigma^2/\delta^2$, a common rule of thumb.

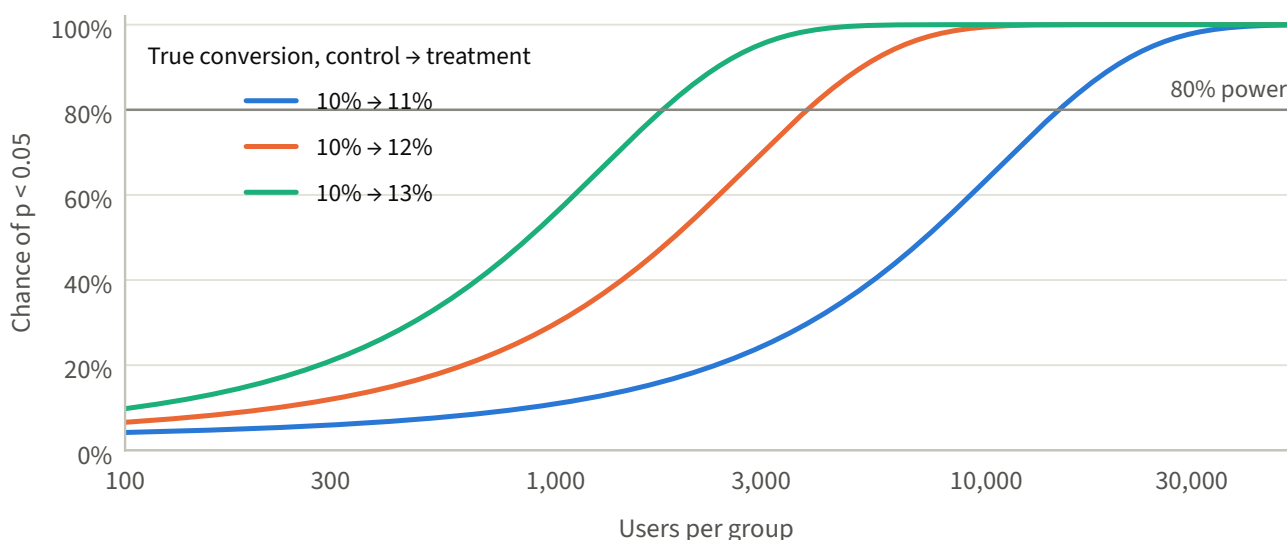
Example 8.9 (A checkout test, by hand)

To detect the lift from 10% to 11% of the previous example with 80% power at level 0.05:

$$\begin{aligned} n &= 7.849 \times (0.09 + 0.0979)/0.01^2 && \text{(the corollary)} \\ &= 1.4748/0.0001 && \text{(multiply)} \\ &= 14,748 \text{ users per group} && \text{(divide, round)} \end{aligned}$$

The rule of thumb gives $16 \times 0.09 / 0.0001 = 14,400$. If 3,000 users reach the checkout each day and half of them enter the test, each group gains $3,000 / 2 / 2 = 750$ users a day, and the test needs $14,748 / 750 = 19.7$ days, so 20; rounding up to 21 days covers each weekday three times. The power formula at other sizes, computed as in the previous example (for 5,000 users: $SE = 0.00613$, the square root of $0.1879 / 5000$, $\Phi(0.01 / 0.00613 - 1.96) = \Phi(-0.33) = 0.37$):

users per group	5,000	10,000	14,748	20,000
power	37%	64%	80%	90%



The power formula against users per group for a 10% baseline and three true treatment rates. A 1-point lift needs 14,748 users per group for 80% power; a 3-point lift needs $7.849 \times (0.09 + 0.13 \times 0.87) / 0.03^2 \approx 1,770$.

Power depends on δ/SE , and SE falls like $1/\sqrt{n}$, so halving the detectable effect needs four times the users.

In ABMeter, the test is fixed horizon and makes no correction for several treatments, so the duration should be chosen in advance with this calculation and the result read once.

Exercises

1. A page converts 5% and the team wants 80% power to detect 6% at level 0.05. How many users per group are needed? At 1,000 new users per group per day, how many days?
2. With the same baseline, the target effect is halved to 5.5%. About how many users per group are needed, by the scaling argument, and exactly?
3. An experiment has four treatments and one control. What per-comparison level does the Bonferroni correction prescribe for a family-wise level of 0.05?

Statistics that are not means

Every interval in this book comes from the central limit theorem applied to means of independent per-user values, possibly passed through a smooth function. Some natural summaries are not of that form, and the theory does not apply to them. This short chapter names the two common ones.

The idea

Everything in this book rests on averaging many small independent contributions, one per user. A sum of per-user numbers has that shape; so does any average built from such sums. A maximum does not: one extreme user decides it. A count of distinct values does not either: two users who bought the same plan count once together, not once each.

Definition 9.1 (Additive statistic)

A group statistic is **additive** over users if it is a sum of one number per user, $\sum_i a_i$; its per-user average is then a sample mean, and Chapters 2, 3 and 6 apply.

Example 9.2 (A maximum and a distinct count)

Four users' largest orders are \$120, \$35, \$0 and \$60, and they bought the plans {pro}, {pro}, none and {basic}.

- The group's largest order is \$120. The per-user values add to 215, so the maximum is not additive: it is decided by one extreme user.
- The group bought 2 distinct plans. The per-user counts of distinct plans add to 3, because two users bought the same plan. Distinct values overlap across users, so a distinct count is not additive either.

In 2,000 simulated samples of lognormal revenue, the two kinds of statistic behave in opposite ways as the sample grows:

users	sd of the mean	sd of the maximum
100	5.52	157.8
1,000	1.74	238.8
10,000	0.56	349.2

The standard deviation of the mean shrinks like $1/\sqrt{n}$, while that of the maximum grows.

Example 9.3 (Total revenue is additive)

The same four users spent \$150, \$35, \$0 and \$90 in total. The group's revenue, \$275, is the sum of the per-user values, so revenue per user, $275 / 4 = 68.75$, is a sample mean, and its standard error is $s/\sqrt{n}$ (the sample standard deviation over $\sqrt{n}$) as in Chapter 1.

Remark 9.4 (What theory applies instead)

The maximum of n observations is described by extreme-value theory, whose limits are not normal, and a distinct count by the theory of occupancy problems. Neither is a mean of independent per-user values, so the standard error $s/\sqrt{n}$, the z-test and the delta method of this book do not apply to them. A count of distinct users is the exception: within one user it is 0 or 1, so it is additive and is the conversion indicator.

In ABMeter, this is why metric formulas accept counts, sums and distinct counts of users, and reject `min`, `max` and distinct counts of other fields.

Exercises

1. Five users bought the plans {basic}, {pro}, {pro, team}, none and {team}. What is the sum of the per-user distinct counts, and how many distinct plans did the group buy?
2. Which of these group statistics are additive over users: (a) the number of distinct users who bought, (b) the number of distinct plans bought, (c) total revenue, (d) the largest order?

Formula sheet

n observations per group, C control, T treatment, Φ the standard normal distribution function, $z_{0.975} = 1.96$.

One sample (Chapters 1 and 2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\text{SE}(\bar{X}) = \sqrt{s^2/n}$$

0/1 data with observed rate $\hat{p}$: $s^2 = n \hat{p}(1 - \hat{p})/(n - 1)$. Variance of a sum: $\text{Var}(X + Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$. Skewness of a mean: $\gamma/\sqrt{n}$.

Two groups (Chapter 3)

$$z = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\text{SE}_T^2 + \text{SE}_C^2}}$$

$p = 2(1 - \Phi(|z|))$; 95% interval $(\bar{X}_T - \bar{X}_C) \pm 1.96 \text{ SE}$; the interval excludes 0 exactly when $p < 0.05$.

Delta method (Chapter 6)

$$\text{SE}^2(f(\bar{m})) = \frac{\mathbf{g}^\top \Sigma \mathbf{g}}{n}$$

$\mathbf{g}$: the gradient of f at the sample means; Σ : the sample covariance matrix of the per-user values. For $R = \bar{X}/\bar{Y}$:

$$\text{SE}_R^2 = \frac{s_X^2 - 2R s_{XY} + R^2 s_Y^2}{n \bar{Y}^2}$$

Bias of R as an estimate of μ_X/μ_Y : about $(\mu_X \sigma_Y^2 - \mu_Y \sigma_{XY})/(n \mu_Y^3)$.

Relative lift (Chapter 7)

$$\text{SE}_{\text{rel}}^2 = \frac{\text{SE}_T^2}{\bar{X}_C^2} + \frac{\bar{X}_T^2 \text{SE}_C^2}{\bar{X}_C^4}$$

Fieller: all θ with $(\bar{X}_T - \theta \bar{X}_C)^2 \leq 1.96^2 (\text{SE}_T^2 + \theta^2 \text{SE}_C^2)$.

Planning (Chapter 8)

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (\sigma_C^2 + \sigma_T^2)}{\delta^2}$$

For $\alpha = 0.05$ and 80% power the factor is 7.849, about $16 \sigma^2/\delta^2$ per group. Power: $\Phi(|\delta|/\text{SE} - 1.96)$. Bonferroni: level α/k for k comparisons.

Hash assignment (Chapter 4)

$$b(s, u) = \left\lfloor \frac{100 \cdot \text{num}(s, u)}{2^{64}} \right\rfloor + 1$$

num: the first 64 bits of SHA-256 of “salt:user”. Two steps need two independent salts.

Glossary

The Russian column gives the usual term in Russian statistical literature.

Term	Meaning	Русский
Random variable	A number whose value depends on chance.	случайная величина
Distribution function	$F(x) = P(X \leq x)$.	функция распределения
Expectation, mean	The probability-weighted average of a random variable, μ .	математическое ожидание
Variance	The expected squared deviation from the mean, σ^2 .	дисперсия
Standard deviation	The square root of the variance.	стандартное отклонение
Covariance, correlation	How two variables move together; correlation is covariance rescaled to $[-1, 1]$.	ковариация, корреляция
Independence	Knowing one variable does not change the distribution of the other.	независимость
i.i.d. sample	Independent draws from one distribution.	независимая одинаково распределённая выборка
Estimator	A function of the sample that approximates an unknown quantity.	оценка
Bias	Expected estimate minus the true value.	смещение
Consistent	Converges to the true value as the sample grows.	состоятельная
Sample variance	s^2 , squared deviations from $\bar{X}$ divided by $n - 1$.	выборочная дисперсия
Sampling distribution	The distribution of an estimator over repeated samples.	выборочное распределение
Standard error	The standard deviation of an estimator.	стандартная ошибка
Law of large numbers	The sample mean converges to the true mean.	закон больших чисел
Normal distribution	The bell-shaped distribution $N(\mu, \sigma^2)$; standard normal when $N(0, 1)$.	нормальное распределение
Quantile	z_q with $\Phi(z_q) = q$, for example $z_{0.975} = 1.96$.	квантиль
Convergence in distribution	Distribution functions converge at every continuity point of the limit.	сходимость по распределению
Central limit theorem	The standardized sample mean is approximately standard normal for large n .	центральная предельная теорема
Skewness	The standardized third central moment; 0 for symmetric distributions.	коэффициент асимметрии
Null hypothesis	The claim tested against, usually “no difference”.	нулевая гипотеза
Test statistic	A number from the data with a known distribution under the null.	статистика критерия
p-value	The probability, under the null, of a statistic at least as extreme as observed.	p-значение
Significance level	The threshold α for rejecting the null.	уровень значимости
Type I, type II error	A false positive; a false negative.	ошибка первого и второго рода
Power	The probability of rejecting the null when an alternative is true.	мощность критерия

Term	Meaning	Русский
Confidence interval	An interval built so that a stated share of such intervals contains the true value.	доверительный интервал
Coverage	The share of intervals that actually contain the true value.	покрытие
Potential outcomes	A unit's outcomes with and without treatment.	потенциальные исходы
Average treatment effect	The mean of individual effects.	средний эффект воздействия
Selection bias	Bias from treatment that depends on the outcome.	смещение отбора
Confounder	A cause of both the treatment and the outcome, such as user type when users choose their version.	конфаундер, вмешивающийся фактор
Seeing and doing	$P(Y X)$ observes users who happen to have X ; $P(Y \text{do}(X))$ asks what setting X does. Randomization makes them equal.	наблюдение и вмешательство
Unit of randomization, of analysis	What the assignment chooses; what a metric averages over.	единица рандомизации, единица анализа
Ratio of means	$\bar{X}/\bar{Y}$, a ratio of totals.	отношение средних
Delta method	The standard error of a smooth function of means from its gradient.	дельта-метод
Gradient	The vector of partial derivatives.	градиент
Forward-mode differentiation	Computing a value and its gradient in one pass.	прямой режим автоматического дифференцирования
Bootstrap	Resampling the data with replacement to estimate a standard error.	бутстрэп
Relative lift	$\mu_T/\mu_C - 1$.	относительный прирост
Fieller's interval	An exact confidence set for a ratio of normal means.	интервал Филлера
Fixed horizon	Sample size fixed in advance, result read once.	фиксированный горизонт
Peeking	Testing repeatedly and stopping at the first significant result.	подглядывание
Bonferroni correction	Testing each of k comparisons at α/k .	поправка Бонферрони

References

1. **Deng, A., Knoblich, U. and Lu, J. (2018).** Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas. *Proceedings of KDD 2018*, 233–242, [doi:10.1145/3219819.3219919](https://doi.org/10.1145/3219819.3219919); preprint [arXiv:1803.06336](https://arxiv.org/abs/1803.06336). The practical reference for Chapters 6 and 7. Section 2 derives the delta-method interval for a percent change, compares it with Fieller’s interval in simulations (Table 1) and proposes a bias correction; Section 3 treats metrics whose analysis unit (page views, orders) differs from the randomization unit (users).
2. **Kohavi, R., Tang, D. and Xu, Y. (2020).** *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press. [doi:10.1017/9781108653985](https://doi.org/10.1017/9781108653985); companion site experimentguide.com. The standard practitioner’s book on online experiments: randomization units, metrics, variance of ratio metrics, sample ratio checks and common pitfalls. Background for Chapters 4, 5 and 8.
3. **Wasserman, L. (2004).** *All of Statistics: A Concise Course in Statistical Inference*. Springer. [doi:10.1007/978-0-387-21736-9](https://doi.org/10.1007/978-0-387-21736-9); author’s page stat.cmu.edu/~larry/all-of-statistics. A compact, rigorous text covering Chapters 1 to 3 and the delta method: random variables, convergence (the law of large numbers, the central limit theorem, Slutsky’s theorem), the bootstrap and hypothesis testing.
4. **Johari, R., Pekelis, L. and Walsh, D. J. (2015).** Always Valid Inference: Bringing Sequential Analysis to A/B Testing. [arXiv:1512.04922](https://arxiv.org/abs/1512.04922). Defines p-values and confidence intervals that remain valid however often the data are inspected and whenever the test is stopped; the remedy for peeking in Chapter 8.
5. **Johari, R., Koomen, P., Pekelis, L. and Walsh, D. (2017).** Peeking at A/B Tests: Why It Matters, and What to Do About It. *Proceedings of KDD 2017*, 1517–1525, [doi:10.1145/3097983.3097992](https://doi.org/10.1145/3097983.3097992). Explains, for practitioners, why continuously monitoring an A/B test inflates false positives and how always-valid inference avoids it.
6. **Fieller, E. C. (1954).** Some Problems in Interval Estimation. *Journal of the Royal Statistical Society, Series B*, 16(2), 175–185. [doi:10.1111/j.2517-6161.1954.tb00159.x](https://doi.org/10.1111/j.2517-6161.1954.tb00159.x). The original construction of the confidence set for a ratio of means used in Chapter 7.
7. **Imbens, G. W. and Rubin, D. B. (2015).** *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press. [doi:10.1017/CBO9781139025751](https://doi.org/10.1017/CBO9781139025751). The potential-outcomes framework of Chapter 4, including randomization inference for completely randomized experiments.
8. **Efron, B. and Tibshirani, R. J. (1993).** *An Introduction to the Bootstrap*. Chapman & Hall/CRC (Springer). [doi:10.1007/978-1-4899-4541-9](https://doi.org/10.1007/978-1-4899-4541-9). The bootstrap standard error and its limits, used as a check in Chapter 6.
9. **Lemire, D. (2018).** Fast Random Integer Generation in an Interval. [arXiv:1805.10941](https://arxiv.org/abs/1805.10941). Analyses the multiply-and-shift mapping of a random integer into a range, used for hash buckets in Chapter 4.
10. **Pearl, J. (2009).** *Causality: Models, Reasoning, and Inference*. 2nd edition, Cambridge University Press. [doi:10.1017/CBO9780511803161](https://doi.org/10.1017/CBO9780511803161)

Answers to the exercises

Every number below is also computed by a script that the book's build runs and checks.

Chapter 1

1. The values are 1, 0, 0, 1, so $n = 4$ and $\bar{X} = 2/4 = 0.5$ (the sample mean). The deviations are ± 0.5 .

$$s^2 = (4 \times 0.25)/3 = 0.3333 \quad (\text{squared deviations over } n - 1)$$

$$SE = \sqrt{0.3333/4} = 0.2887 \quad (SE = s/\sqrt{n})$$

$$s^2 = 4 \times 0.5 \times 0.5/3 = 0.3333 \quad (\text{Bernoulli shortcut})$$

2. Apply the variance of a sum to $X + (-Y)$:

$$\begin{aligned} s_{X-Y}^2 &= s_X^2 + s_Y^2 - 2s_{XY} && (\text{variance of a sum, with } -Y \text{ in place of } Y) \\ &= 320 + 1.7 - 46 = 275.7 && (\text{substitute}) \end{aligned}$$

3. The standard error is proportional to $1/\sqrt{n}$, so halving it needs four times the users: $4 \times 400 = 1,600$.

Chapter 2

1. With $\sigma = 22.98$ (the standard deviation of one user's value): $22.98 / \sqrt{100} = 2.298$, and $22.98 / \sqrt{10,000} = 0.2298$.

2. The skewness of the mean of 30 users is $10.6 / \sqrt{30} \approx 1.94$. For skewness 0.1:

$$10.6/\sqrt{n} = 0.1 \quad (\text{skewness of the mean, set to the target})$$

$$\sqrt{n} = 106 \quad (\text{solve})$$

$$n = 106^2 = 11,236 \quad (\text{square})$$

3. Chebyshev's bound and the central limit theorem:

$$\sigma^2/(n\varepsilon^2) = 23^2/(10,000 \times 1^2) = 0.0529 \quad (\text{law of large numbers bound})$$

$$SE = 23/\sqrt{10,000} = 0.23 \quad (\sigma/\sqrt{n})$$

$$\varepsilon/SE = 1/0.23 = 4.35 \quad (\text{the distance in standard errors})$$

By the central limit theorem the probability is about $2(1 - \Phi(4.35)) \approx 1.4 \times 10^{-5}$. Chebyshev's bound holds for any distribution and is far from tight.

4. With $\sigma = 20$ and $n = 400$:

$$SE = 20/\sqrt{400} = 1 \quad (\sigma/\sqrt{n})$$

$$P(|\bar{X} - \mu| > 2) = 2(1 - \Phi(2)) \quad (2 \text{ is two standard errors; two tails})$$

$$= 2 \times (1 - 0.9772) = 0.0456 \quad (\text{table value})$$

The exact Φ gives 0.0455. For the band:

$$1.96 \times 20/\sqrt{n} = 0.5 \quad (\text{half-width of the 95\% band, set to 0.5})$$

$$\sqrt{n} = 1.96 \times 20/0.5 = 78.4 \quad (\text{solve})$$

$$n = 78.4^2 = 6,146.56 \quad (\text{square})$$

rounded up to 6,147.

Chapter 3

1. With the observed rates 0.10 and 0.125 and $n = 1,000$ per group:

$$\begin{aligned} SE &= \sqrt{(0.1 \times 0.9 + 0.125 \times 0.875)/999} && \text{(0/1 shortcut)} \\ &= \sqrt{0.199375/999} = 0.014127 && \text{(evaluate)} \\ z &= 0.025/0.014127 = 1.77 && \text{(definition of } z\text{)} \\ p &= 2 \times (1 - 0.9616) = 0.077 && \text{(two-sided p-value)} \\ \text{interval} &= 0.025 \pm 1.96 \times 0.014127 = [-0.0027, 0.0527] && \text{(interval)} \end{aligned}$$

The script, with the exact Φ , gives $p = 0.0768$. Not significant at 0.05: $p > 0.05$ and the interval contains 0.

2. (b) is true by definition, and (c) is true by the duality of tests and intervals. (a) is false: a p-value is computed assuming H_0 (the null hypothesis) and is not a probability of H_0 . (d) is false: a p-value does not measure size.
3. No test is significant with probability 0.95^{20} , so the answer is $1 - 0.95^{20} = 0.6415$.

Chapter 4

1. The individual effects are 1, 0, 0, 1, so the ATE (the average treatment effect) is $(1 + 0 + 0 + 1)/4 = 0.5$. Treating users 1 and 3 compares their $Y(1)$ with the $Y(0)$ of users 2 and 4:

$$\begin{aligned} \bar{Y}_T - \bar{Y}_C &= (3 + 1)/2 - (0 + 1)/2 && \text{(treated users 1, 3; control users 2, 4)} \\ &= 2 - 0.5 = 1.5 && \text{(divide and subtract)} \end{aligned}$$

By the randomization proposition the average over all 6 assignments is the ATE, 0.5.

2. Control shrinks from buckets 1–50 to 1–40, so buckets 41–50 move to treatment: 10 of every 100 test users, 10%.
3. Independent salts: $0.5 \times 0.3 = 0.15$. The same salt: buckets 1–30 are inside 1–50, so 30% of users are in both.

Chapter 5

1. Mean of ratios: $(40/2 + 50/1) / 2 = 35$, from the 2 users with orders. Ratio of means: $(40 + 50) / (2 + 1) = 30$, from all 4 users.
2. Revenue per user doubles (factor 2). Average order value does not change (factor 1): revenue and orders both double.
3. Let the m users with $Y_i > 0$ all have $Y_i = c$, and let the sums below run over those users.

$$\begin{aligned} M &= \frac{1}{m} \sum \frac{X_i}{c} = \frac{\sum X_i}{mc} && \text{(definition of the mean of ratios)} \\ R &= \frac{\sum_{\text{all}} X_i}{\sum_{\text{all}} Y_i} = \frac{\sum X_i + 0}{mc + 0} && \text{(users with } Y_i = 0 \text{ have } X_i = 0) \end{aligned}$$

The two fractions are equal, so $M = R$.

Chapter 6

1. The gradient is (2), so $SE^2 = 2^2 s^2/n = 4s^2/n$ and the standard error doubles: a constant factor scales the value and its standard error alike.
2. The value is $4 \times 3 = 12$. By the product rule the gradient is $(\bar{V}, \bar{C}) = (3, 4)$, and

$$\begin{aligned} g^T \Sigma g &= 3^2 s_C^2 + 2 \times 3 \times 4 s_{CV} + 4^2 s_V^2 && \text{(quadratic form for } k = 2) \\ &= 9 s_C^2 + 24 s_{CV} + 16 s_V^2 && \text{(multiply)} \end{aligned}$$

3. With $s_{XY} = -80$ the middle term changes sign:

$$\begin{aligned} s_X^2 - 2R s_{XY} + R^2 s_Y^2 &= 1610 + 2 \times 19 \times 80 + 361 \times 8 && \text{(substitute } s_{XY} = -80) \\ &= 1610 + 3040 + 2888 = 7538 && \text{(multiply and add)} \\ SE_R &= \sqrt{7538/10} = 27.46 && \text{(divide by } n \bar{Y}^2 = 10) \end{aligned}$$

With a negative covariance the errors of revenue and orders reinforce each other in the ratio instead of cancelling.

Chapter 7

1. With $\bar{X}_T = 0.08$, $\bar{X}_C = 0.10$ and both standard errors 0.01:

$$\begin{aligned} \hat{L} &= 0.08/0.10 - 1 = -0.20 && \text{(relative lift)} \\ SE_{\text{rel}} &= \sqrt{0.0001/0.01 + 0.0064 \times 0.0001/0.0001} && \text{(delta method)} \\ &= \sqrt{0.0164} = 0.1281 && \text{(add, root)} \\ \text{delta interval} &= -0.20 \pm 1.96 \times 0.1281 = [-0.451, 0.051] && \text{(delta method)} \\ SE_{\text{naive}} &= \sqrt{0.0001 + 0.0001/0.10} = 0.1414 && \text{(naive)} \\ \text{naive interval} &= -0.20 \pm 1.96 \times 0.1414 = [-0.477, 0.077] && \text{(naive)} \end{aligned}$$

The naive interval is wider, because the treatment is below control and the delta method multiplies the control term by $(0.08/0.10)^2 = 0.64$.

- If $\bar{X}_T = \bar{X}_C$, then $\bar{X}_T^2 SE_C^2 / \bar{X}_C^4 = SE_C^2 / \bar{X}_C^2$, and the two variance formulas are equal.
- No. Fieller's set is a finite interval only when $\bar{X}_C^2 > 1.96^2 SE_C^2$; here $100 < 1.96^2 \times 36 = 138.3$, so the control mean is not significantly different from 0 and the confidence set for the ratio is unbounded.

Chapter 8

1. The per-user variances are $0.05 \times 0.95 = 0.0475$ and $0.06 \times 0.94 = 0.0564$.

$$\begin{aligned} n &= 7.849 \times (0.0475 + 0.0564)/0.01^2 && \text{(the sample-size corollary)} \\ &= 8.155/0.0001 = 8,155 && \text{(multiply and divide)} \end{aligned}$$

At 1,000 users per group per day, $8,155 / 1,000 = 8.2$ days, so 9 days, or 14 to cover each weekday twice.

- Halving the effect multiplies n by 4: about $4 \times 8,155 \approx 32,600$. The exact formula, with the treatment variance $0.055 \times 0.945 = 0.051975$ at 5.5%, gives 31,231.
- $0.05 / 4 = 0.0125$ per comparison.

Chapter 9

- The per-user distinct counts are 1, 1, 2, 0 and 1, summing to 5. The group bought 3 distinct plans: basic, pro and team.
- (a) and (c). A user's count of distinct users is 0 or 1 and a user's revenue is one number, so both group values are sums over users. Plans overlap across users, and the largest order is decided by one user.