Introduction: The Population is the Point

PSCI1801 Chapter 1

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

What Statistics Is Actually About

The point of statistics is to make **inferences** *about a population using a sample of data.*

Describing the sample in front of you is not *the goal. The goal is what your sample tells you about the population.*

What defines statistics: making conclusions about populations from a sample.

What "Population" Means

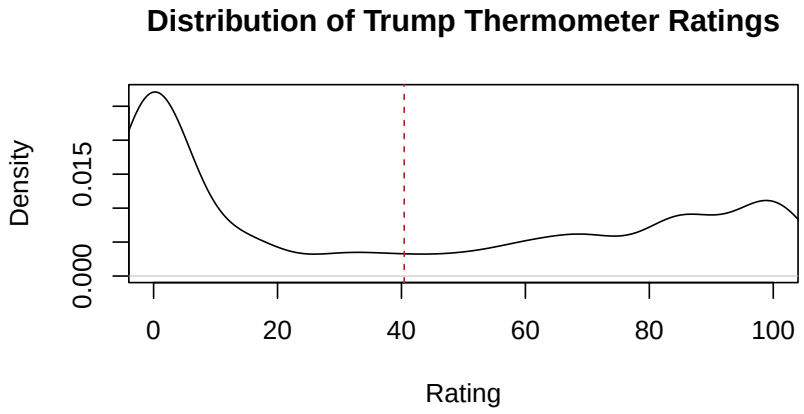
A vague term covering a wide range of things:

- **Obvious case**: the adult U.S. public, sampled by a poll of 8,000.
- **Business earning potential**: a few months of finances = a sample from the business's true earning distribution.
- **Baseball hitting ability**: a stretch of at-bats = a sample from the player's true hitting ability.
- **Data-generating process (DGP)**: the underlying process producing your data, whether or not a well-defined "population" exists.

A DGP produces data probabilistically, *not deterministically. We will never know the truth exactly — we form estimates and quantify our uncertainty about them.*

A Motivating Example

2020 ANES: ~8,000 Americans rate Trump on a 0–100 "feeling thermometer."



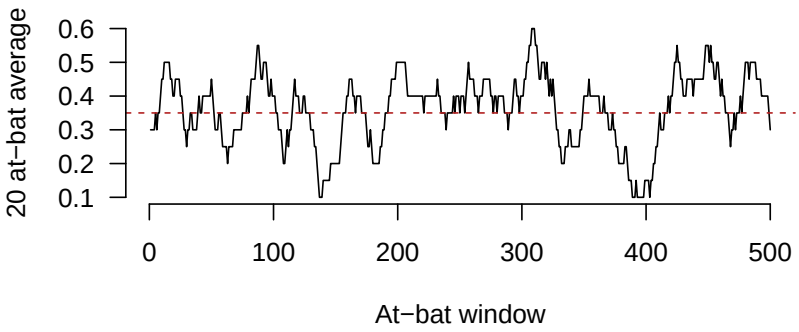

```
 mean    sd
40.44 40.31
```

Lots of density near 0 and near 100; very few ambivalent responses. But this describes our sample. *What we care about is how* all *voters feel.*

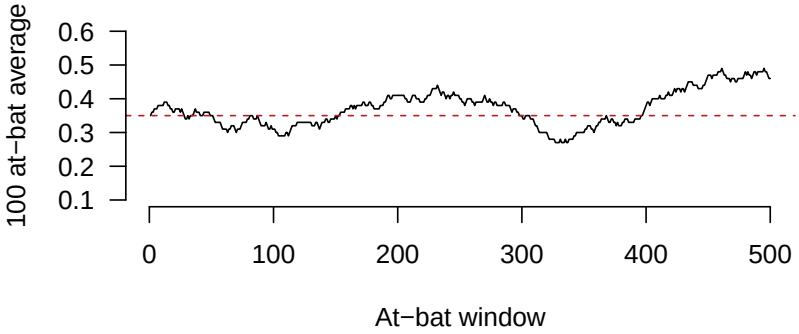
If we ran the survey again, would we get exactly 40.44? Obviously not. So: how much could the mean vary? Could we get 45? 60? Is 42 more likely than 50 — and by how much?

Why the Sample Isn't Automatically the Answer

Suppose a batter's true ability is a .350 average. Simulate 10,000 at-bats and look at rolling 20-at-bat windows (about 4 games).



The truth is fixed at .350, yet 4-game stretches swing between .100 and .600 by pure chance. *If our rule is "believe whatever the sample says," we will be wrong constantly.*

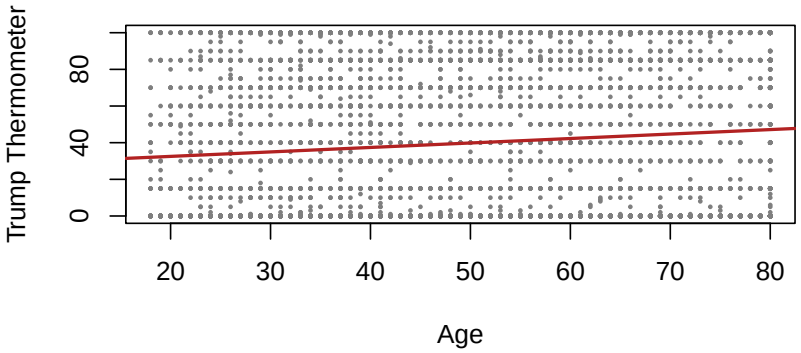


A larger sample gives a smoother line closer to the truth — but it still ranges from about .300 to .500. Randomness alone produces outcomes far from the truth.

We need a way to measure and express how much confidence to place in what a sample tells us.

From Describing to Inferring

Real data: does age predict feelings toward Trump?



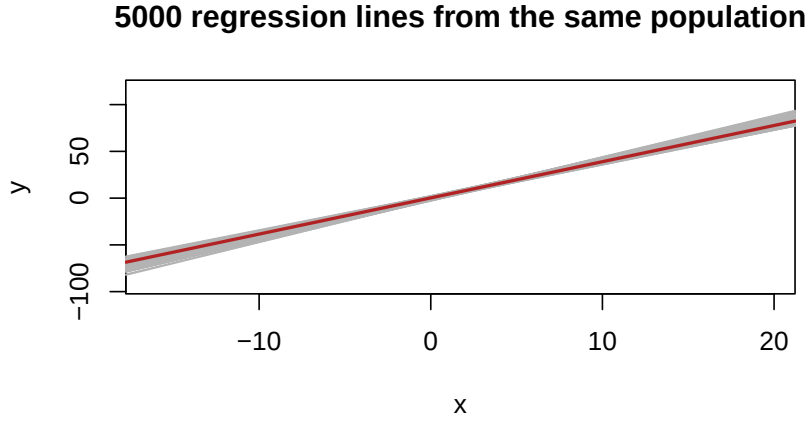
	Est	SE	t	p
(Intercept)	27.646	1.440	19.20	<0.001
age	0.244	0.027	9.16	<0.001

Intercept 27.64: predicted rating at age 0 (impossible, but mechanically required). Slope 0.24: each additional year of age is associated with a 0.24-point warmer rating in this sample.

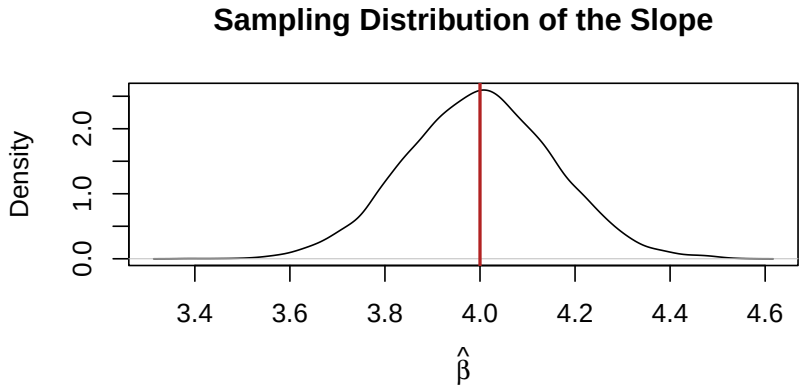
But we care about the relationship in the population. *The* `Std. Error`, `t value`, *and* `p` *columns are what let us get there.*

The Magic: Sampling Distributions

Build fake data where we know *the truth — the real slope is 4 — then re-sample 5,000 times and record the slope each time.*



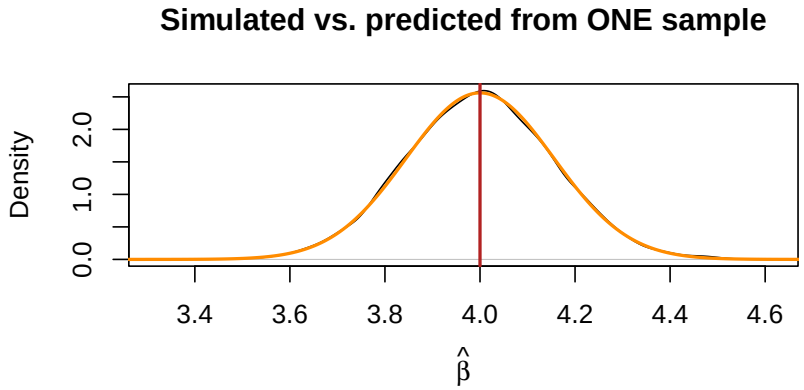
Every sample gives a slightly different line. What does the distribution of those 5,000 slopes look like?



It is a **bell curve** — *specifically a* **normal distribution**, *centered on the truth.*

The Standard Error Contains All of It

Here is the genuinely remarkable part. We cannot re-sample 5,000 times in real life. But the `Std. Error` *from* one *regression tells us the width of that entire distribution.*



SE from one regression	SD of 5000 slopes
0.1557	0.1551

The black curve took 5,000 samples. The orange curve took one. *They are the same distribution.*

The Main Question of the Course

> *"If there is nothing going on — if there is* no *relationship in the population — how likely is it to observe the pattern I see in my sample?"*

Every hypothesis test, confidence interval, and p-value is a different formalization of that one question.

If the answer is "very unlikely," we conclude something real is happening. If "not unlikely," we cannot rule out that our result is noise.

Takeaway: *the goal of statistics is to understand the population, not to describe the sample. We imagine calculating our statistic many times; the resulting distribution is a knowable, calculable thing, and that is what lets us judge how special our result is.*

Probability

PSCI1801 Chapter 3

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Why Probability at All?

Motivating case (Mlodinow, The Drunkard's Walk*): Bill Miller beat the S&P 500 for 15 straight years. One analyst put the odds at 372,529:1. Is that the right calculation?*

The framing question for the whole course: **if this were pure chance, how likely is what I observed?** *If it happens easily by chance, we should not be impressed.*

Later we ask the same question about regression: "if there were no *relationship in the population, how likely is a sample relationship this extreme?"*

The Basics

An **event** *is an outcome. The* **sample space** (Ω) *is the set of all possible outcomes.*

- *Dice:* $\{1, 2, 3, 4, 5, 6\}$ *Coin:* $\{H, T\}$

Probability is the **expected long-run frequency** *of an event:*

> *The chance of something gives the percentage of time it is expected to happen, when the basic process is done over and over again, independently and under the same conditions.*

When all outcomes are equally likely, *probability is just counting:*

$$P(\text{event}) = \frac{\#\text{ outcomes in the event}}{\#\text{ outcomes in } \Omega}$$

So $P(even) = 3/6 = 1/2$. *This does* **not** *work when outcomes are unequally likely —* $P(third\ party\ wins) \neq 1/3$.

The axioms

1. Non-negativity. $P(A) \geq 0$

2. Something happens. $P(\Omega) = 1$ *(so probabilities live in* $[0, 1]$*).*

3. Addition rule. *If* A *and* B *are* mutually exclusive *(one occurring rules out the other):*

$$P(A \text{ or } B) = P(A) + P(B)$$

But be careful — "roll a 4 or an even number" are not *mutually exclusive. Naively adding gives* $1/6 + 1/2 = 4/6$*, which is wrong; counting the cells gives* $1/2$*. The general form subtracts the double-count:*

$$P(A \text{ or } B) = P(A) + P(B) - P(A \,\&\, B)$$

This also covers the mutually exclusive case, where $P(A\&B) = 0$*.*

Complements

The **complement** A^c *is everything in* Ω *that is not* A*. Since an event either happens or it doesn't:*

$$P(A) + P(A^c) = 1 \quad \Longrightarrow \quad P(A) = 1 - P(A^c)$$

This turns out to be enormously useful — often the "not" version is far easier to count.

Simulating Probability in R

Because probability is long-run frequency, we can estimate *it by doing the thing many times. Throughout this course: do it by math, then confirm by simulation.*

The engine is the `for()` *loop. Odds of 0 heads in 7 flips:*

```r
set.seed(19104)
coin <- c(0,1)                    #0 = tails, 1 = heads
#empty vector to store results
num.heads <- rep(NA, 100000)
for(i in 1:100000){
 num.heads[i] <-  sum(sample(coin,7,replace=T))
}
prop.table(table(num.heads))
```

```
num.heads
      0       1       2       3       4       5       6
0.00787 0.05551 0.16051 0.27450 0.27651 0.16201 0.05540
      7
0.00769
```

Note the three-part pattern you will use all semester: **(1)** *create an empty vector,* **(2)** *loop, writing one result into position* `i` *each pass,* **(3)** *summarize the vector.* R runs a loop silently — *if you don't save the result, nothing is recorded.*

Permutations & Combinations

A **permutation** *is a sequence where order* matters. *A* **combination** *is one where it does not.*

$$ {}_nP_k = \frac{n!}{(n-k)!} \qquad {}_nC_k = \frac{n!}{k!\,(n-k)!} $$

The combination formula is the permutation formula divided by $k!$ *— collapsing all orderings of the same* k *items into one.*

Locker problem. *50 numbers, 3 unique numbers in order:* $50 \times 49 \times 48 = 117{,}600$ *permutations, so* $P(\text{guess}) = 1/117{,}600$.

Mega Millions. *5 white balls from 70 (order irrelevant) plus 1 gold ball from 25:*

```r
choose(70,5)
```

```
[1] 12103014
```

When is the lottery a good bet? Expected value = benefits − costs:

$$E[\text{Lottery}] = P(\text{win}) \times \text{Prize} - \text{Cost}$$

Breaking even at $2 requires a prize over $605 million. Which happens! So should you always play then? No — as the jackpot grows, more people play, *raising the chance you must* split *the prize.*

The birthday problem

In a class of 20, what is the probability two people share a birthday? Use the complement — count the sequences with no *shared birthday:*

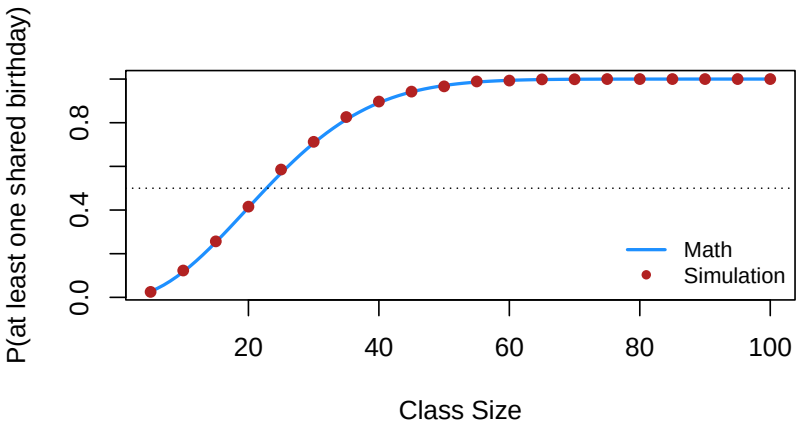
$$P(\text{no shared}) = \frac{{}_{365}P_{20}}{365^{20}} = \frac{365!/(365-20)!}{365^{20}}$$

R cannot compute 365! *directly (the number is too large), so build it from* `choose()`*:*

```
(choose(365,20)*factorial(20))/365^20
```

```
[1] 0.5885616
```

So $P(\text{at least one shared}) \approx 41\%$ — strikingly high. Math and simulation agree across all class sizes:



Conditional Probability

Two dice: the sample space is $6 \times 6 = 36$ *equally likely cells. Each cell shows the sum.*

	1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	**8**
3	4	5	6	7	**8**	**9**
4	5	6	7	**8**	**9**	**10**
5	6	7	**8**	**9**	**10**	**11**
6	7	**8**	**9**	**10**	**11**	**12**

Counting the 15 bold cells: $P(\text{sum} \geq 8) = 15/36 = 5/12 \approx 0.42$.

Now suppose I tell you die 1 came up 5. We are no longer working with 36 outcomes — only the six in that row:

	1	**2**	**3**	**4**	**5**	**6**
5	6	7	**8**	**9**	**10**	**11**

$$P(\text{sum} \geq 8 \mid \text{die 1} = 5) = \frac{4}{6} = \frac{2}{3}$$

The key insight: *conditioning is a* **subsetting** *operation. It changes the sample space, which changes the probability. The general formula:*

$$P(A \mid B) = \frac{P(A \,\&\, B)}{P(B)}$$

Dividing by $P(B)$ *is what re-scales the surviving outcomes so they sum to 1 again.*

$$P(A \mid B) \neq P(B \mid A)$$

These are different questions. We just found $P(\text{sum} \geq 8 \mid \text{die 1} = 5) = 2/3$. *Reversing it, restrict to the 15 cells with sum* ≥ 8 *and count those with die* $1 = 5$ *— there are 4:*

$$P(\text{die } 1 = 5 \mid \text{sum} \geq 8) = \frac{4}{15} \approx 0.27$$

Very different. Confusing these two is one of the most common and most consequential errors in applied statistics.

Independence

Two events are **independent** *when knowing one tells you nothing about the other:*

$$P(A \mid B) = P(A) \qquad \text{and} \qquad P(B \mid A) = P(B)$$

Roll a die and flip a coin — 12 equally likely outcomes:

$$P(2 \mid H) = \frac{P(H \,\&\, 2)}{P(H)} = \frac{1/12}{6/12} = \frac{1}{6} = P(2) = \frac{2}{12}$$

Equal, so they are independent — as they obviously should be. This concept becomes critical later: we assume our samples are **iid** *(independent and identically distributed).*

The Monty Hall Problem

> *Three doors: one car, two goats. You pick one. Monty opens a different door revealing a goat, then offers you the switch. Should you?*

People imagine six equally likely outcomes. The real sample space has only three, because **Monty never opens the car door** *— his choice is not random:*

Stay	Switch
Car $\to$ Car	Car $\to$ Goat
$\text{Goat}_1 \to \text{Goat}_1$	$\text{Goat}_1 \to$ Car
$\text{Goat}_2 \to \text{Goat}_2$	$\text{Goat}_2 \to$ Car

The cleanest way to see it: $P(\text{Car} \mid \text{picked a goat first \& switch}) = 1$. *You pick a goat first with probability* $2/3$*, so switching wins* $2/3$ *of the time.*

$$P(\text{Car} \mid \text{Stay}) = \tfrac{1}{3} \qquad P(\text{Car} \mid \text{Switch}) = \tfrac{2}{3}$$

```
stay <- NA
switch <- NA
for(i in 1:10000){
  pick <- sample(c("goat","goat","car"),1)

  stay[i] = pick

  if(pick=="car"){
    switch[i] = "goat"
  } else {
    switch[i]="car"
  }
}
mean(stay=="car")
```

```
[1] 0.3318
```

```
mean(switch=="car")
```

```
[1] 0.6682
```

This also introduces `if () ... else ...` inside a loop — run different code depending on a condition.

Resolving the Bill Miller Case

The wrong question: *what are the odds* this specific person *beats the market for* these specific 15 years?

```
set.seed(19104)
coin <- c(0,1)

all.heads <- rep(NA, 1000000)

for(i in 1:length(all.heads)){
  s <- sample(coin, 15, replace=T)
  all.heads[i] <- sum(s)==15
}

sum(all.heads)
```

```
[1] 33
```

About 33 in a million — genuinely rare.

The right question: *with ~6,000 advisers working over ~40 years, what are the odds that* someone *has a 15-year streak? The* `rle()` *function finds run lengths:*

```
adv.streak <- function(n.adv, n.yrs, streak){
  str.res <- NA
  for(i in 1:n.adv){
    s <- sample(coin, n.yrs, replace=T)
    str.res[i] <- sort(rle(s)$lengths[rle(s)$values==1],
                       decreasing = T)[1]>=streak
  }
  return(any(str.res))
}

any.lucky <- NA
for(i in 1:100){
any.lucky[i] <- adv.streak(n.adv=6000, n.yrs = 40,
                           streak=15)
}
table(any.lucky)
```

```
any.lucky
FALSE  TRUE
    6    94
```

In roughly 94 of 100 simulated worlds, at least one *adviser gets a 15-year streak by luck alone. It would be strange if* nobody *did.*

Change the parameters and the conclusion flips. A boutique firm with 5 traders, one of whom beat the market all 10 years it has existed:

```
any.lucky <- NA
for(i in 1:1000){
any.lucky[i] <- adv.streak(n.adv=5, n.yrs = 10, streak=10)
}
table(any.lucky)
```

```
any.lucky
FALSE  TRUE
  989     6
```

About 2% by luck alone — now I might genuinely have a special employee.

The lesson: *the "impressiveness" of a result depends entirely on how many chances there were for it to happen. Getting the reference class right is the whole game.*

Coming Next

We now have machinery for the probability of specific events. Next we step up in abstraction: instead of computing "the probability of 15 heads," we describe the underlying randomness as a **random variable** *— a data-generating process with parameters (expected value, variance) that summarize the whole process without simulating every outcome.*

Discrete Random Variables

PSCI1801 Chapter 4

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

The Motivating Problem

Patrick Mahomes goes 0/3 to start the Superbowl. Is he having a bad night, or is this ordinary chance?

With 3 flips of a fair *coin we can just write out the sample space — 8 equally likely rows:*

$$p(K=k) = \begin{cases} 1/8 & \text{if } 0 \\ 3/8 & \text{if } 1 \\ 3/8 & \text{if } 2 \\ 1/8 & \text{if } 3 \end{cases}$$

So 0/3 happens 12.5% of the time by chance — not alarming.

But two things break this approach:

1. *What about 0 for 10? We cannot write out 1,024 rows.*
2. *Mahomes completes 67% of passes, not 50%. Counting only works when outcomes are* **equally likely**.

We need a general formula. That formula is a **random variable**.

Deriving the Binomial

Step 1: how many ways?

How many ways to get k heads in n flips? Order does not matter — "successes in slots 1 and 2" is the same sequence as "slots 2 and 1" — so this is a **combination**, *not a permutation:*

$${}_nC_k = \frac{n!}{k!\,(n-k)!}$$

Two heads in three flips: ${}_3C_2 = 3$. *Two heads in ten flips:* ${}_{10}C_2 = 45$.

For a fair *coin that is enough: divide by the* 2^n *total sequences.* $P(2 \text{ heads in } 3) = 3/8$; $P(2 \text{ in } 10) = 45/1024 \approx 0.04$.

Step 2: how likely is each way?

For an unfair *coin the sequences are no longer equally likely. Use the multiplication rule for independent events,* $P(A\&B) = P(A)P(B)$*:*

Sequence	Probability
H (.67) H (.67) T (.33)	0.148
H (.67) T (.33) H (.67)	0.148
T (.33) H (.67) H (.67)	0.148

Each of the 3 ways has probability $0.67^2 \times 0.33 = 0.148$*, so* $P(2 \text{ heads}) = 3 \times 0.148 = 0.444$.

In general, any sequence with k *successes has probability* $\pi^k(1-\pi)^{n-k}$.

The binomial formula

Multiply (number of ways) by (probability of each way):

$$\boxed{P(K = k) = {}_nC_k \, \pi^k \, (1-\pi)^{n-k}}$$

```
prob.k <- function(n,k,p){
  choose(n,k)*p^(k) * (1-p)^(n-k)
}
# 2 completions in 3 attempts
round(prob.k(3, 2, 0.67), 4)
```

```
[1] 0.4444
```

Math and simulation agree — as they should:

```r
set.seed(19104)
coin <- c(rep(1,67), rep(0,33))

two.heads <- rep(NA, 100000)

for(i in 1:length(two.heads)){
  flip <- sample(coin, 3, replace = T)
  two.heads[i] <- sum(flip)==2
}

mean(two.heads)
```

```
[1] 0.44548
```

Formally Defining a Random Variable

A **random variable** *assigns a number and a probability to every possible outcome of a repeatable process. Because outcomes now have* numbers, *we can do math on them.*

Vocabulary warning. *Three terms mean essentially the same object in this course:*

random variable = population = data generating process (DGP)

Values must be **mutually exclusive and exhaustive** *— every outcome gets exactly one value.*

Two flavors: **discrete** *(finite/countable values: coin flips, passes completed) and* **continuous** *(any real value in an interval: height, GDP — next chapter).*

Two named RVs

- **Bernoulli**: *a single trial.* $P(X = 1) = \pi$, $P(X = 0) = 1 - \pi$.
- **Binomial**: *n iid Bernoulli trials, counting successes.*

iid = *independent and identically distributed. Independent: one flip doesn't change the next. Identically distributed: same probability every time. A random-sample survey is iid; so is a roulette wheel (regardless of the casino's display of past spins).*

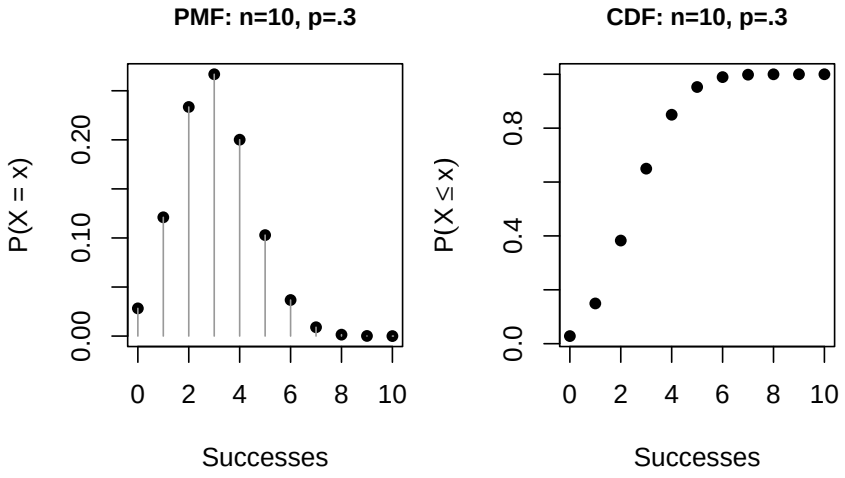
PMF and CDF

The **probability mass function (PMF)** *gives the probability of each value:* $f(x) = P(X = x)$.

The **cumulative distribution function (CDF)** *gives the probability of that value* or less:

$$F(x) = P(X \leq x) = \sum_{k \leq x} f(k)$$

The CDF starts at 0, never decreases, and ends at 1. PMF and CDF are 1:1 *— give me one and I can give you the other.*



Notice the first CDF value equals the first PMF value, and the second CDF value is the first two PMF values added together.

R gives us these directly: `dbinom()` *for the PMF,* `pbinom()` *for the CDF,* `rbinom()` *to draw samples.*

```
round(dbinom(0:5, size = 10, prob = .3), 4)   # PMF
```

```
[1] 0.0282 0.1211 0.2335 0.2668 0.2001 0.1029
```

```
round(pbinom(0:5, size = 10, prob = .3), 4)   # CDF
```

```
[1] 0.0282 0.1493 0.3828 0.6496 0.8497 0.9527
```

The CDF answers genuinely useful questions. If completions were random at 66.5%, how often would a QB complete 25 or more of 35 passes?

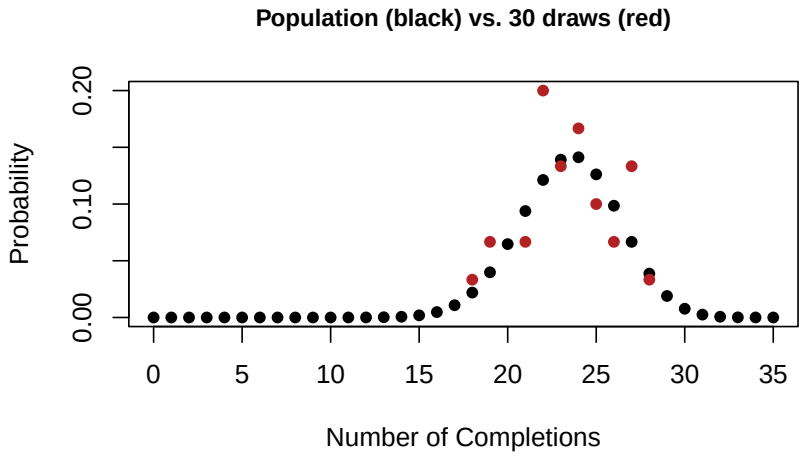
```
round(1 - pbinom(24, 35, .665), 4)   # P(25 or more)
```

```
[1] 0.3367
```

Samples From a Random Variable

This is the most important idea in the chapter, and the bridge to sampling.

The PMF says what should *happen in the long run. But we never flip a coin infinitely — we get* one realization.



The red dots were generated by *the black dots, and are clearly related to them — but they are not identical. Just because we are* supposed *to get 24 completions 15% of the time does not mean we will.*

The gap between what the RV says should happen and what a finite sample actually produces is exactly where this course is headed.

Expectation and Variance

In a sample *we describe data with the mean and standard deviation. For a* population/RV *the equivalents are the* **expected value** *and* **variance**. *These describe distributions, not samples.*

$$E[X] = \sum_x x\,p(x) \qquad V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

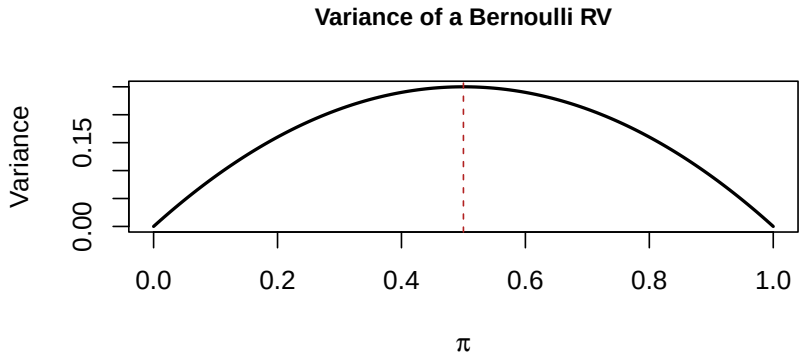
The expected value is the average value we would get sampling a large number of times — each outcome weighted by its probability.

Bernoulli

$$E[X] = 0(1-\pi) + 1(\pi) = \pi \qquad V[X] = \pi - \pi^2 = \pi(1-\pi)$$

(Using $1^2 = 1$ *and* $0^2 = 0$*, so* $E[X^2] = E[X] = \pi$*.)*

A neat result: Bernoulli variance is largest at $\pi = 0.5$ *and shrinks toward the extremes. A near-certain process has little variability.*



Binomial

The long way, for $n = 3$, $\pi = 0.5$*:*

$$E[K] = 0\tfrac{1}{8} + 1\tfrac{3}{8} + 2\tfrac{3}{8} + 3\tfrac{1}{8} = 1.5$$

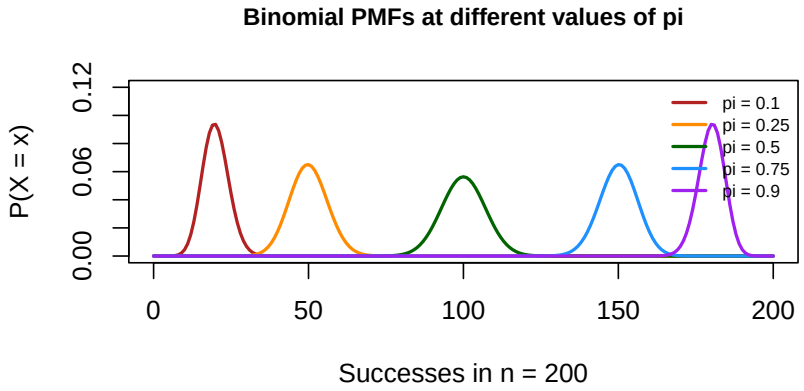
$$E[K^2] = 3, \quad E[K]^2 = 2.25 \quad \Rightarrow \quad V[K] = 3 - 2.25 = 0.75$$

The shortcut — because a Binomial is just n iid Bernoullis:

$$\boxed{E[X] = n\pi \qquad V[X] = n\pi(1-\pi)}$$

Check: $E = 3(0.5) = 1.5$ *and* $V = 3(0.5)(0.5) = 0.75$. *Same answers, far less work.*

Variance scales linearly *with n and is again maximized at $\pi = 0.5$.*



Note how the distributions are widest near $\pi = 0.5$ and tighten at the extremes — which affects how easily we can spot an outlier under different assumed truths.

Coming Next

Every RV here has been discrete. *Real measurements are often* continuous *— and, more importantly, the sampling distributions we are building toward are continuous too. That vocabulary is the last piece before sampling.*

Continuous Random Variables

PSCI1801 Chapter 5

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Why We Need Continuous RVs

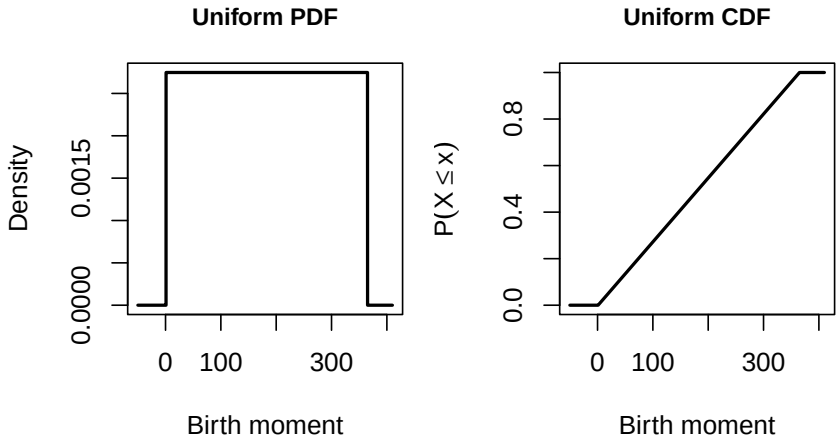
Bernoulli and Binomial are **discrete** *— a countable number of outcomes. But height, income, temperature, and GDP can take* any *real value in an interval.*

The deeper reason: next chapter, the **sampling distribution** *of a statistic is itself a random variable — and a continuous, normally distributed one. We need this vocabulary first.*

The Uniform Random Variable

All points in an interval are equally likely. Running example: the exact moment *in the year you are born, from 0 to 365.*

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$



The critical idea: P(exact value) = 0

There are infinitely many *possible birth moments. If every specific moment had positive probability, the total would be infinite — not 1.*

So for any continuous RV, the probability of any exact value is zero. *You can be born at day 5.182191710…, and the probability of that precise instant is 0.*

Because of this the graph is no longer a probability mass *function. For continuous RVs it is a probability* **density** *function (PDF). R will still return a number from* `dunif()` *or* `dnorm()` *— that is the* height of the curve, *needed to draw it, not a probability.*

PDF and CDF, in calculus

For discrete RVs we summed. *For continuous RVs we* integrate:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \qquad\qquad f(x) = \frac{d}{dx}F(x)$$

The CDF is the area under the PDF; the PDF is the slope *of the CDF. They remain 1:1 — give me one, I give you the other.*

Expectation and variance

$$E[X] = \frac{a+b}{2} \qquad\qquad V[X] = \frac{1}{12}(b-a)^2$$

For birth moment: $E[X] = 365/2 = 182.5$ *(July 1, noon) and* $V[X] = \frac{1}{12}(365)^2 = 11{,}102$*, so* $\sigma = 105.4$.

(The 12 has nothing to do with months — it is in the formula for every uniform RV.)

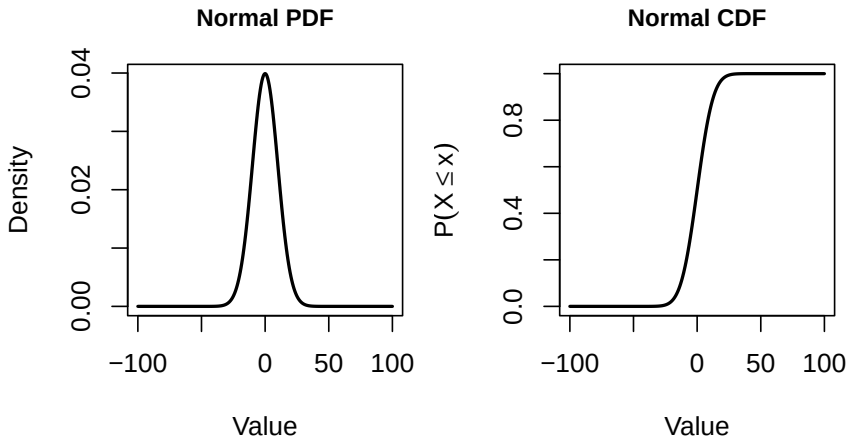
The Normal Random Variable

The famous one. Also called Gaussian.

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

You never need to reproduce this. The only *thing that matters: the sole unknowns are x, μ, and σ. So* **if you know μ and σ you can draw the entire distribution.**

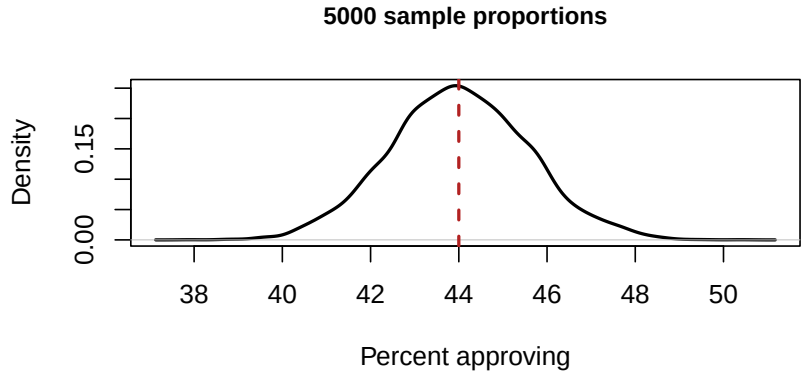
It runs from $-\infty$ to $+\infty$ and is perfectly symmetric *about μ — which means μ is both the mean and the median, with 50% of the mass on each side.*



Why the normal matters so much

(1) *Many natural quantities are normally distributed (e.g. height).*

(2) *Far more important: the* **sampling distribution** *of nearly every statistic we care about is normal. Take 5,000 surveys of 1,000 people where the truth is 44% approval:*



Normal — and this holds no matter what the starting population looks like. That is where the course is headed.

The Standard Normal

Unlike the binomial (where variance depends on π), for a normal the **variance is unrelated to the mean**. *μ sets location, σ sets shape. Two rules follow:*

$$Z = X + c \;\Rightarrow\; Z \sim N(\mu + c,\ \sigma^2)$$

$$Z = cX \;\Rightarrow\; Z \sim N(c\mu,\ c^2\sigma^2)$$

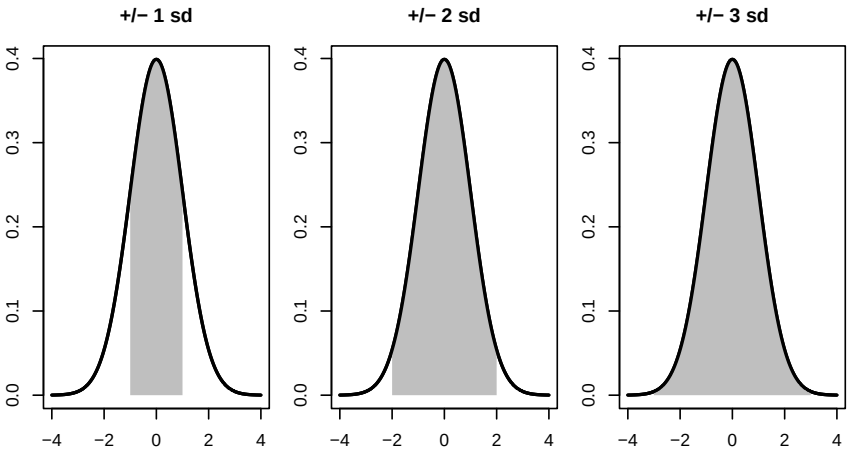
Adding a constant slides the distribution; multiplying rescales it. Together they let us transform any *normal into one specific normal:*

$$\boxed{Z = \frac{X - \mu}{\sigma} \sim N(0, 1)}$$

Subtract the mean, divide by the standard deviation, and you always get the **standard normal**: *mean 0, variance 1. (`dnorm()` and friends default to exactly this.)*

The empirical rule

In the standard normal, one "unit" is one standard deviation. Because every normal converts to the standard normal, the following holds for every *normal distribution:*



```r
1-2*pnorm(-1)
```

```
[1] 0.6826895
```

```r
1-2*pnorm(-2)
```

```
[1] 0.9544997
```

```r
1-2*pnorm(-3)
```

```
[1] 0.9973002
```

$$\mu \pm 1\sigma \approx 68.3\% \qquad \mu \pm 2\sigma \approx 95\% \qquad \mu \pm 3\sigma \approx 99.7\%$$

Note the trick used above: because the normal is symmetric, the middle area is $1 - 2 \times P(\text{left tail})$. Using `1-2*pnorm(1)` instead would give a negative number — impossible for a probability. Use your head.

Z-scores

A **Z-score** *says how many σ's a value sits from μ. For women's height, $N(\mu = 64, \sigma = 3)$, a height of 73 has $Z = (73 - 64)/3 = 3$.*

```r
c(via.z.score = pnorm(3),
  directly = pnorm(73, mean = 64, sd = 3))
```

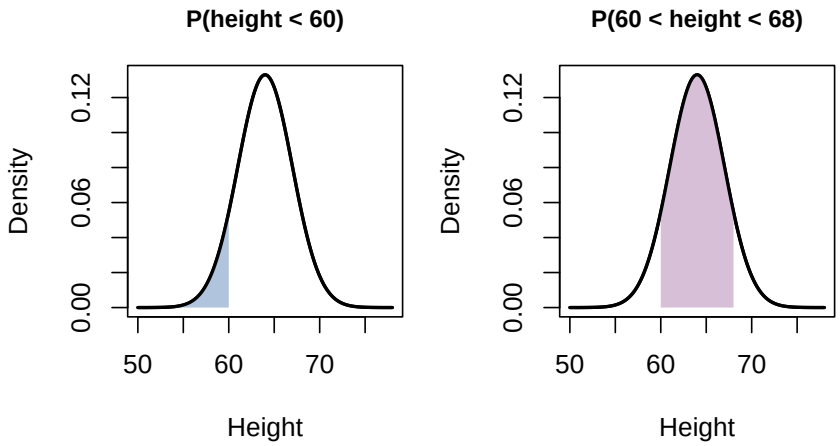
```
via.z.score    directly
  0.9986501   0.9986501
```

Identical — and you should understand exactly why before moving on. Z-scores matter because (a) you will not have R on the hand-written exams, and (b) they give a common language for every normally distributed quantity. The `t value` in regression output is a close cousin: how many standard errors an estimate sits from zero.

Answering Questions With These Distributions

Two classes of question, and one R function for each:

- ***pnorm():*** *given a* value, *what proportion lies to its left (or right)?*
- ***qnorm():*** *given a* proportion, *what value puts that much in the tail?*




```
pnorm(60, 64, 3)  # left of 60
```

```
[1] 0.09121122
```

```
pnorm(60, 64, 3, lower.tail = FALSE)  # right of 60
```

```
[1] 0.9087888
```

```
pnorm(68, 64, 3) - pnorm(60, 64, 3)  # between
```

```
[1] 0.8175776
```

```
qnorm(.8, 64, 3)       # taller than 80% of women
```

```
[1] 66.52486
```

```
qnorm(.025, 64, 3); qnorm(.975, 64, 3)  # middle 95%
```

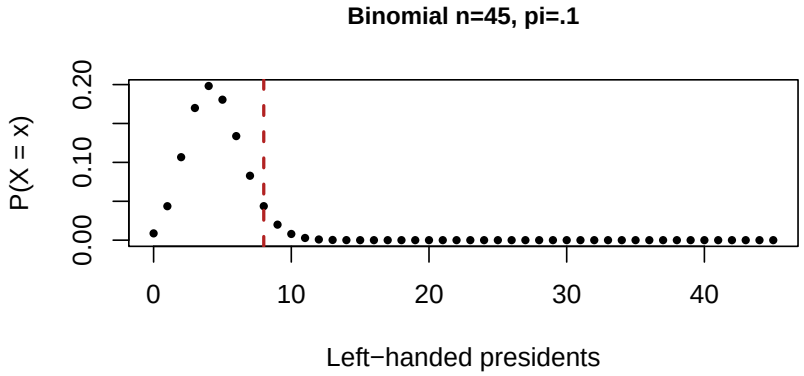
```
[1] 58.12011
```

```
[1] 69.87989
```

`lower.tail = FALSE` flips the direction automatically instead of computing $1 - p$.

A discrete-vs-continuous trap

8 of 45 presidents have been left-handed, against a 10% population base rate. Model as Binomial($n = 45$, $\pi = 0.1$)*, where* $E[X] = n\pi = 4.5$.



We want "at least 8," not "exactly 8." For a discrete *RV you must be careful:*

```
dbinom(8, 45, .1)        # exactly 8
```

```
[1] 0.04370462
```

```
# 8 or more  -- note the 7, not the 8!
1 - pbinom(7, 45, .1)
```

```
[1] 0.07569864
```

Subtracting `pbinom(8, ...)` would wrongly remove the probability of exactly 8. For a continuous *RV this adjustment is unnecessary, since* $P(\text{exact value}) = 0$*:*

```
# born in the last 15 days of the year
1 - punif(350, 0, 365)
```

```
[1] 0.04109589
```

Postwar, 6 of 14 presidents were left-handed:

```
round(1 - pbinom(5, 14, .1), 5)
```

```
[1] 0.00147
```

Under 1% by chance. Seriously — what is going on there?

One more warning: ± 1 *sd contains 68% only for the* normal. *For our uniform birth-moment RV:*

```
round(1 - 2 * punif(182.5 - 105.36, 0, 365), 4)
```

```
[1] 0.5773
```

About 58%, not 68%. The empirical rule is a normal-distribution *fact.*

Coming Next

We now have the vocabulary for discrete and continuous random variables — the language of populations and DGPs. Next we flip the question: what happens when we take a **sample** *from a random variable, and can we work backwards from that sample to statements about the DGP that produced it? That is where probability theory finally becomes statistics.*

Sampling

PSCI1801 Chapter 6

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

The Problem

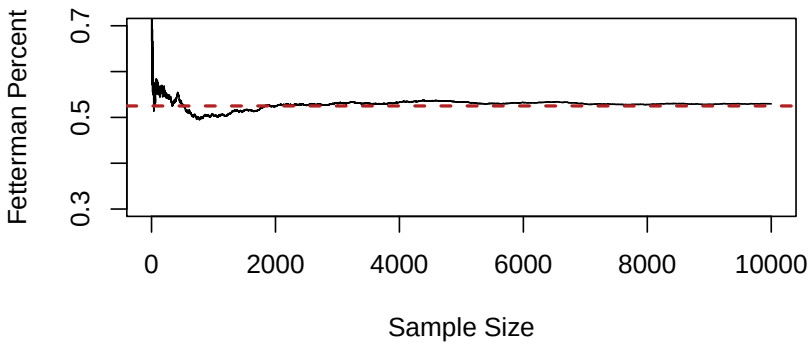
With a known *distribution we can judge how strange an outcome is. Is 6/14 left-handed presidents weird, given a 10% base rate? Is a* -20% *S&P year weird, given* $N(7.5, 18.5^2)$*? In both cases we had a distribution to compare against.*

Now the real case. A 2022 poll of ~5,000 Pennsylvanians gives Fetterman 55.6% two-party support. Suppose the race were actually tied — would 55.6% be crazy?

Crazy compared to what? *There is no obvious distribution to compare against. Some unknown "black box" random variable takes a population truth and produces sample means. We need to understand that black box.*

The Law of Large Numbers

Sample voters one at a time from a Bernoulli$(\pi = .525)$ *and track the running mean.*



At $n = 1$ the mean must be 0 or 1. By $n \approx 2000$ we are close to the truth and stay there.

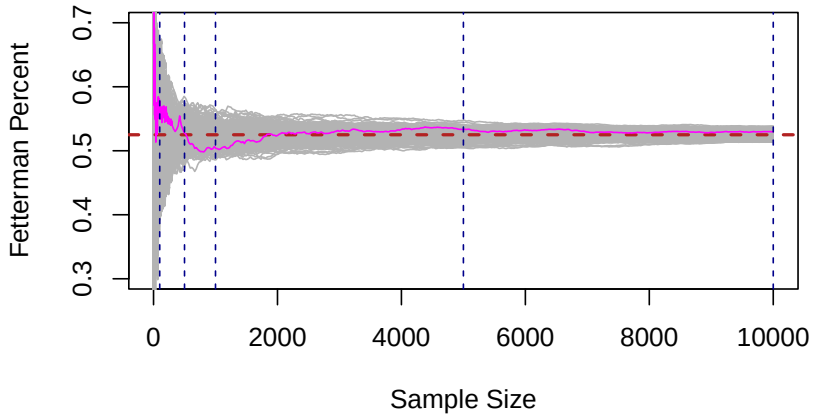
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \;\longrightarrow\; E[X]$$

Principle 1. *The larger the sample, the more likely the sample mean equals the expected value of the RV we are sampling from.*

Notation: X is the RV; X_i is observation i; n is the sample size; the bar always means an average.

From One Run to a Thousand

Now run 1,000 independent samples, each growing from $n = 1$ to $n = 10{,}000$.

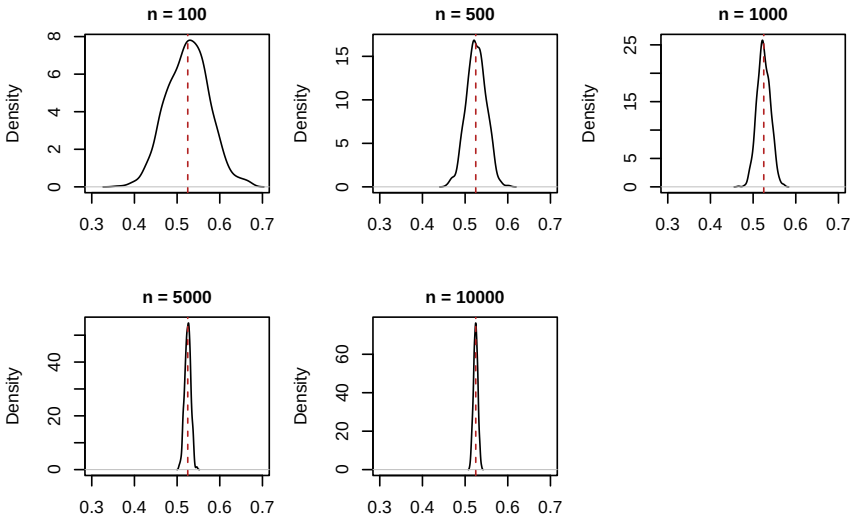


Any vertical cross-section *is 1,000 surveys at that n. Two things to notice: the spread* narrows *as n grows, and at every n the estimates are* centered on the truth.

Principle 2. *For* any *sample size, the expected value of the sample mean is the expected value of the RV we are sampling from.*

Even with $n = 10$ my best guess before sampling is still $E[X]$ — that answer never changes.

Slicing the cross-sections



Every cross-section is a **normal distribution** *centered on the truth. This is wild, and it is the heart of the course.*

A sampling distribution *is the distribution of* all possible estimates *for a given sample size. It is* **not** *the distribution of one sample, and* **not** *the distribution of the population.*

Principle 3. *The sampling distribution is normal, centered on $E[X]$, and gets tighter as n grows.*

We are now tracking two *random variables: X (the Bernoulli producing individual voters) and $\bar{X}$ (the normal producing sample means).*

The Math

Can we know the width of a sampling distribution without *simulating? Yes — derive $E[\bar{X}]$ and $V[\bar{X}]$.*

Expected value

Using $E[cX] = cE[X]$ and $E[X+Z] = E[X]+E[Z]$, and that $E[X_i] = E[X]$ for any single draw:

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum X_i\right] = \frac{1}{n}(E[X] + \cdots + E[X]) = \frac{1}{n}\, n\, E[X] = E[X]$$

Variance

The variance operator behaves differently: $V[cX] = c^2V[X]$, *and* $V[X+Z] = V[X]+V[Z]$ only if independent — *which iid sampling guarantees.*

$$V[\bar{X}_n] = \frac{1}{n^2}(V[X]+\cdots+V[X]) = \frac{1}{n^2}\,n\,V[X] = \frac{V[X]}{n}$$

Check against simulation. Our Bernoulli has $V[X] = \pi(1-\pi) = .525 \times .475 = .249$*:*

```
ns <- c(100, 1000, 10000)
round(rbind(simulated  = apply(binomial.sample[ns, ],
                               1, var),
            calculated = .249 / ns), 6)
```

```
               [,1]     [,2]     [,3]
simulated  0.002449 0.000242 0.000026
calculated 0.002490 0.000249 0.000025
```

The Central Limit Theorem

$$\boxed{\bar{X}_n \rightsquigarrow N\left(\mu = E[X],\ \ \sigma^2 = \frac{V[X]}{n}\right)}$$

Equivalently, in standardized form:

$$\frac{\bar{X}_n - E[X]}{\sqrt{V[X]/n}} \rightsquigarrow N(0,1)$$

The standard deviation of the sampling distribution has its own name — the **standard error**:

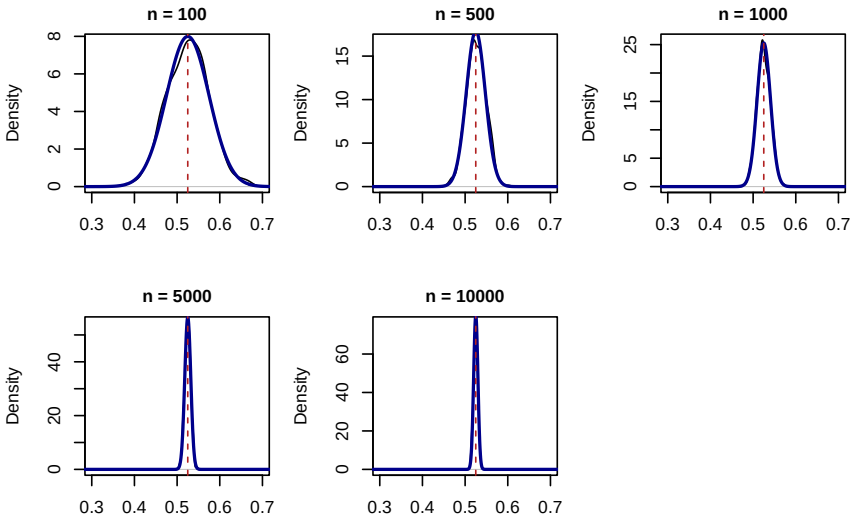
$$SE = \sqrt{V[\bar{X}]} = \sqrt{\frac{V[X]}{n}} = \frac{sd(X)}{\sqrt{n}}$$

The standard deviation of the sampling distribution is the standard error. *(I have a tattoo of this on my arm, it is so important.)*

Think about how good that equation is: less variance in the population → tighter sampling distribution; larger n *→ tighter sampling*

distribution, at a decreasing rate *(the $\sqrt{n}$). Quadruple n to halve the width.*

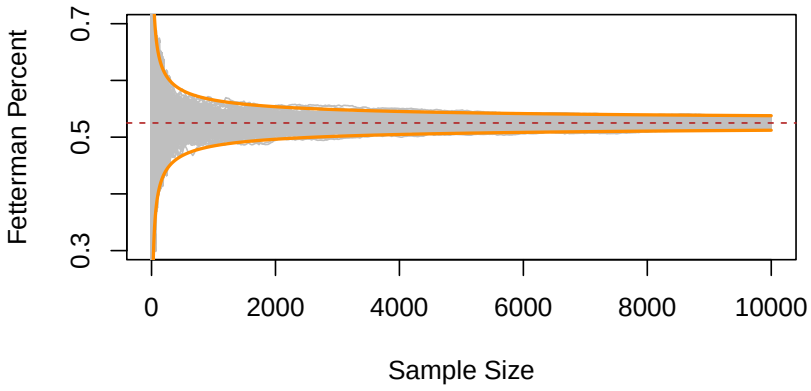
Overlaying the predicted normal on the simulated cross-sections:



Predicting the range before sampling

99% of a normal lies within ± 2.57 SDs (`qnorm(.005)`). So before collecting anything:

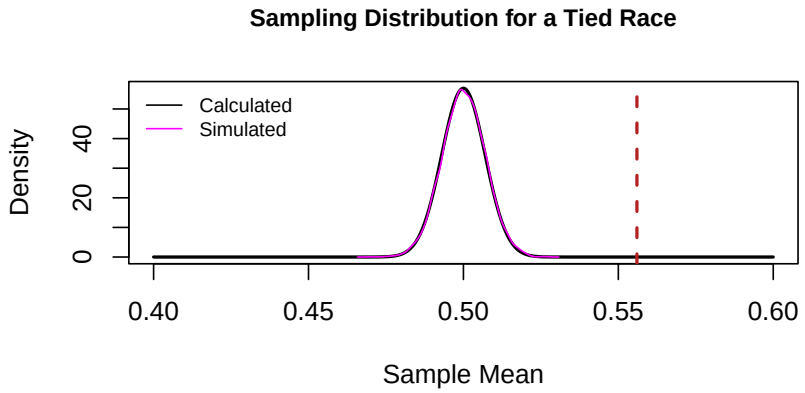
$$.525 \pm 2.57 \times \frac{.499}{\sqrt{n}}$$



Those bounds, derived before any sampling, contain essentially all the samples.

Back to the Poll

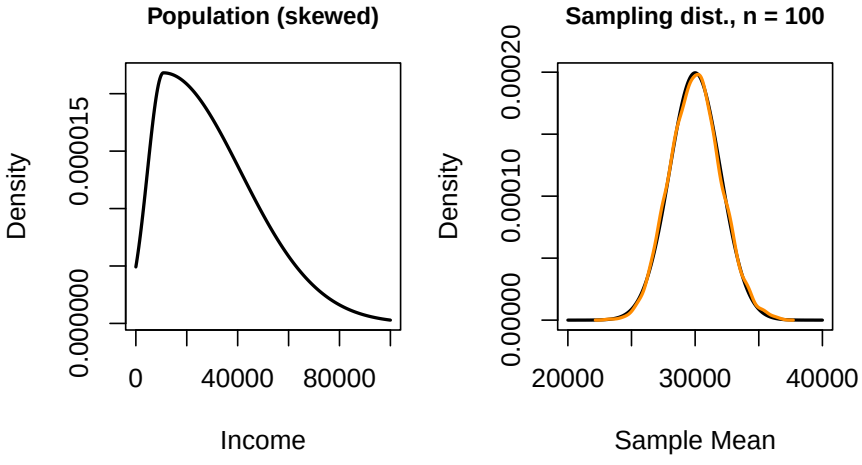
If the race were tied: $E[X] = .5$, $sd(X) = \sqrt{.5 \times .5} = .5$, *and* $SE = .5/\sqrt{5000} = .007$. *So* $\bar{X} \sim N(.5,\ .007^2)$.



Our observed 55.6% is many standard errors away. We would reject the idea that this sample came from a tied race.

It Works for Any Population

Critically, none of this required the population to be normal. Take a badly skewed income distribution with $\mu = 30{,}000$, $\sigma = 20{,}000$*:*



The population is skewed. Every sample is skewed. The sampling distribution is normal. Always.

What Makes a Good Estimator?

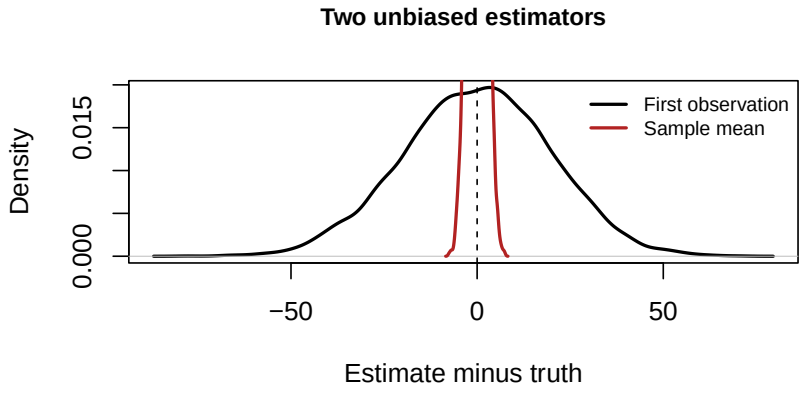
Language: **population parameters** *(fixed, true, unknown) are targeted by* **estimators** *(procedures), which produce* **estimates** *(one number from one sample).*

Unbiasedness

$$E[\hat{\theta} - \theta] = 0$$

The sample mean is unbiased: $E[\bar{X}_n] - E[X] = E[X] - E[X] = 0$.

But so is a silly alternative — "just use the first observation," X_1 *— since* $E[X_1] = E[X]$. *Unbiasedness alone is clearly not enough.*



Both centered on zero; wildly different spreads.

Consistency

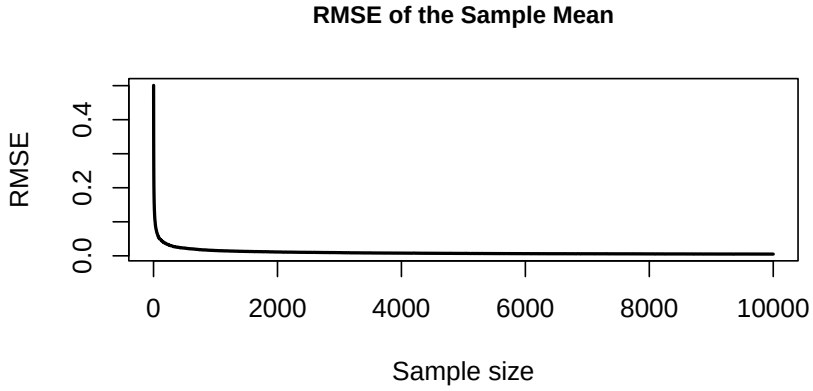
$$V[\bar{X}_n] = \frac{V[X]}{n} \qquad \text{vs} \qquad V[X_1] = V[X]$$

The sample mean's variance shrinks with n; *the first-observation estimator's never does. What we really want is an estimator that gets* better with more data.

Combine both concerns in the **Mean Squared Error**:

$$MSE = V[\hat{\theta}] + \left(E[\hat{\theta} - \theta]\right)^2 = \text{Variance} + \text{Bias}^2$$

A **consistent** *estimator has* $MSE \to 0$ *as* $n \to \infty$. *RMSE (its square root) is easier to read, being on the scale of the variable.*



Tying it together: *because the sample mean is unbiased, the bias term is zero and* MSE *reduces to* $V[X]/n$ *— whose square root is* $sd(X)/\sqrt{n}$, exactly the standard error. *The* $1/\sqrt{n}$ *decay in this curve is the same* $1/\sqrt{n}$ *that set the width of every sampling distribution above. It is all one thing.*

Standard Error of the Mean

PSCI1801 Chapter 7

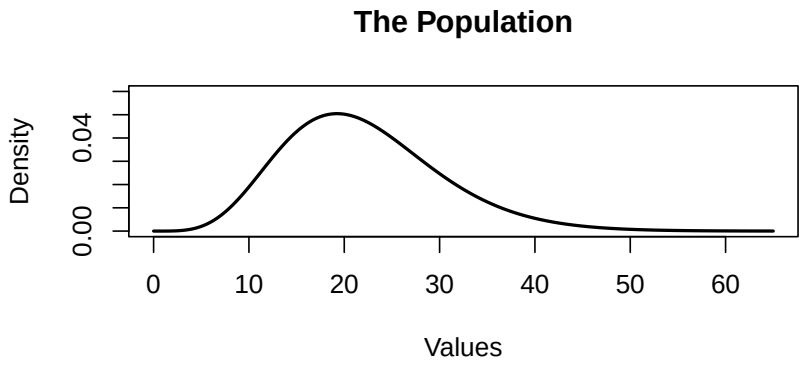
This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

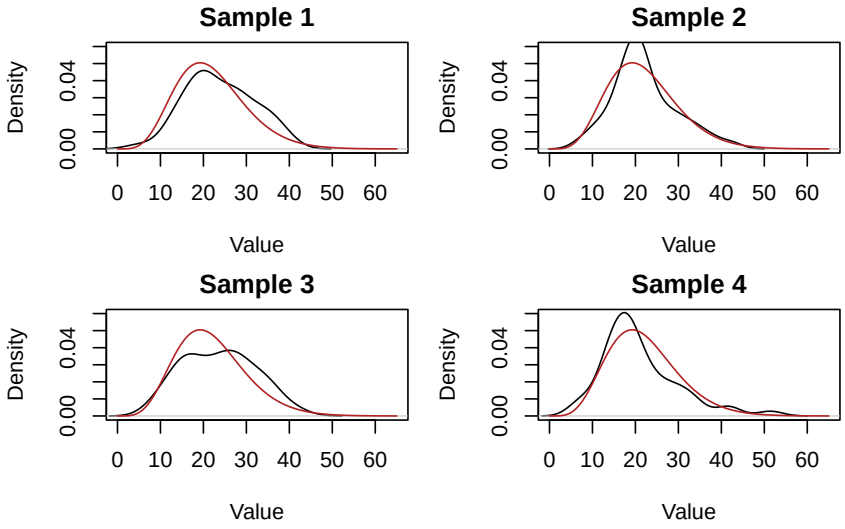
The Three Distributions

Every inferential problem involves three *distinct distributions. Conflating them is the single most common mistake at this stage.*

1. **The population** (the DGP we sample from). *Can be any shape at all.*
2. **The sample.** *The n numbers we actually drew — a rough approximation of the population, different every time.*
3. **The sampling distribution of the sample mean.** *The distribution of $\bar{X}$ we would get by repeating the sampling process. Always (approximately) normal.*

To make the point, let the population be a Chi-squared distribution — deliberately non-normal. Here $E[X] = 22$ and $V[X] = 68$.





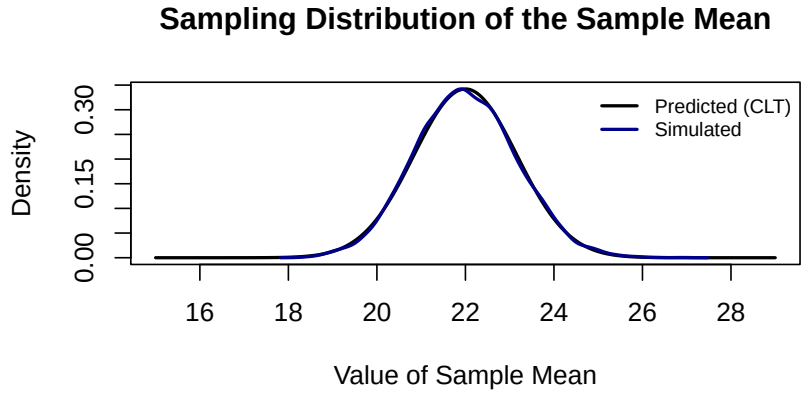
Each sample (black) roughly approximates the population (red) — but only roughly, and never identically.

The key asymmetry: *statistics has* no *good way to answer "how much will my sample's distribution differ from the population's?" But it has a* precise *answer to "how much will my sample* mean *differ from the population mean?"*

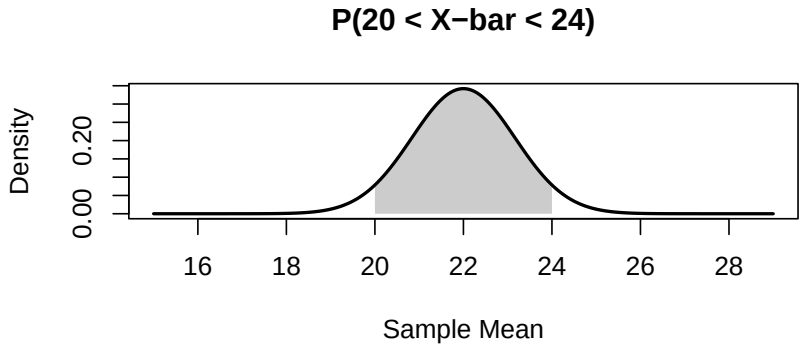
The Sampling Distribution of the Mean

$$\bar{X}_n \sim N\left(\mu = E[X],\ \sigma^2 = \frac{V[X]}{n}\right)$$

With $E[X] = 22$, $V[X] = 68$, $n = 50$, *we can draw this distribution* before *sampling — and confirm it by simulation.*



Identical. Because we know the shape exactly, we can ask precise probability questions — e.g. what share of sample means fall between 20 and 24?




```
theoretical   simulated
     0.9137      0.9194
```

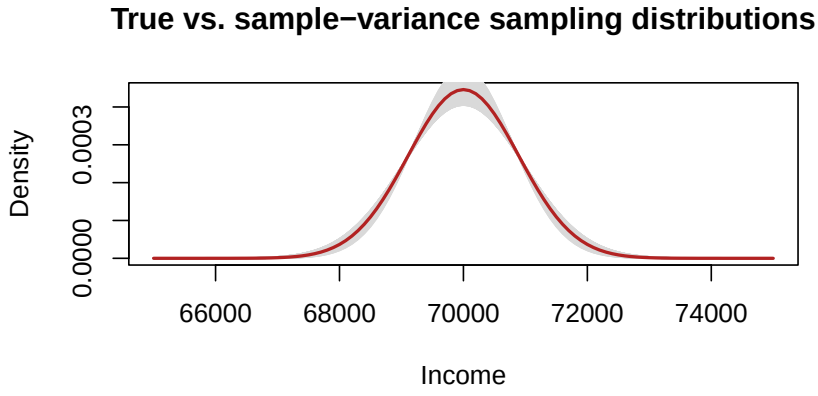
The Real World: We Don't Know the Population

The whole problem: the formula needs $E[X]$ and $V[X]$, and we have neither.

Step 1 — the center. *Change the question. Instead of "what is $E[X]$?", ask "* what if $E[X]$ *were some specific value?" Then the sampling distribution is centered there by assumption.*

Step 2 — the width. *We need $V[X]$. Our only option is to substitute the* sample *variance s^2 as a stand-in for the population variance.*

Does that work? Below: the true sampling distribution (red) against 1,000 sampling distributions built from sample variances (gray).



Close enough to work. Note the sample variance divides by $n-1$, not n:

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

Why? Because only this version is unbiased:

$$E\left[\frac{1}{n-1}\sum(X_i-\bar{X}_n)^2\right] = \sigma^2 \qquad E\left[\frac{1}{n}\sum(X_i-\bar{X}_n)^2\right] \neq \sigma^2$$

`var()` and `sd()` make this correction automatically.

Confidence Intervals

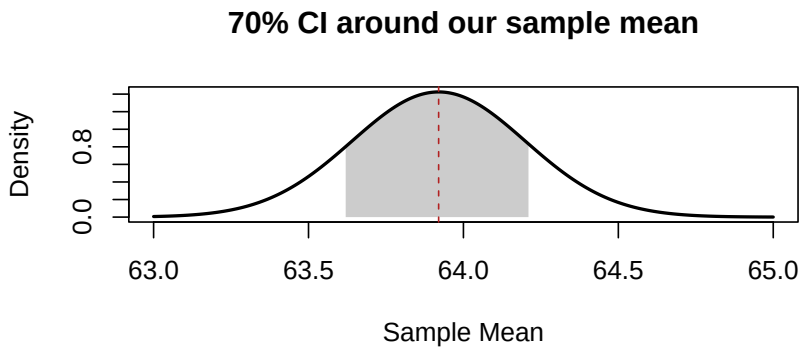
A confidence interval is a zone we are some percent certain contains the true population parameter.

Center the estimated sampling distribution on our sample mean, *then take the region holding $(1-\alpha)\cdot 100$ percent of the probability mass:*

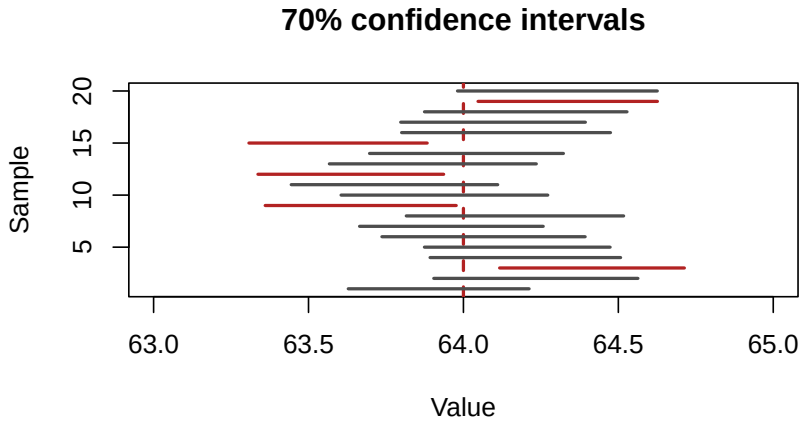
$$CI(\alpha) = \left[\bar{X}_n - z_{\alpha/2}\cdot SE,\ \ \bar{X}_n + z_{\alpha/2}\cdot SE\right] \qquad SE = \frac{s}{\sqrt{n}}$$

Worked example: $n = 100$ from a population with true mean 64. Sample gives $\bar{X} = 63.92$, $s^2 = 7.88$, so $SE = \sqrt{7.88/100} = 0.28$. For 70% confidence, $\alpha = 0.3$ and $z_{\alpha/2} = 1.04$:

$$CI = [63.92 - 1.04(0.28),\ 63.92 + 1.04(0.28)] = [63.62,\ 64.21]$$



Does it work? Draw 20 samples, build a CI from each, and see how many capture the truth.




```
share of CIs containing truth
                        0.6977
```

A note on language. *Strictly, a given CI either does or does not contain the truth — the correct phrasing is "70% of similarly constructed intervals contain the true parameter." Saying "there is a 70% chance this interval contains the truth" is technically wrong, and intuitive enough that we will live with it.*

What makes intervals wider or narrower?

- *Smaller α (more confidence)* $\rightarrow$ **wider**. *To be surer, you need more room.*
- *Larger s (noisier data)* $\rightarrow$ **wider**.
- *Larger n* $\rightarrow$ **narrower** *(the LLN at work).*

The t Distribution

One critical complication. Repeat the coverage check with $n = 10$ instead of 100:

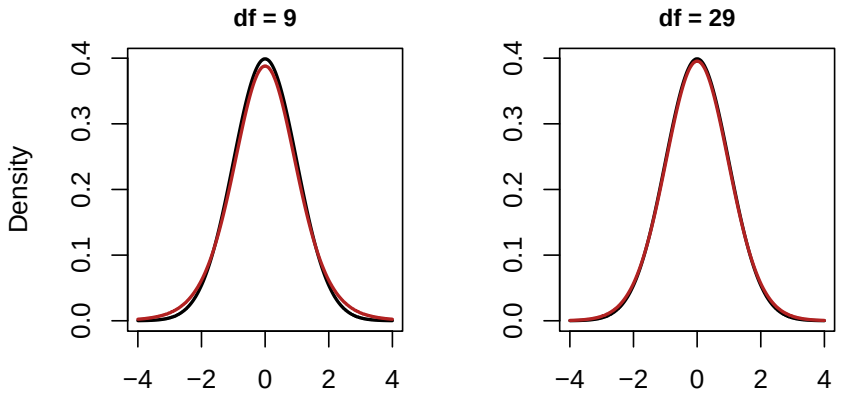
```
coverage using z = 1.04
                 0.682
```

We asked for 70% and got about 68%. Systematically wrong.

The reason: *when we substitute the sample variance for the population variance, the sampling distribution is* no longer normal:

$$\frac{\bar{X}_n - E[X]}{\sqrt{V[X]/n}} \sim N(0,1) \qquad \text{but} \qquad \frac{\bar{X}_n - E[X]}{s/\sqrt{n}} \sim t_{n-1}$$

The t distribution is shorter with fatter tails, and its shape depends on n through its degrees of freedom.




```
normal  t.df9
-1.036 -1.100
```

Using the correct t cutoff fixes the coverage:

```
coverage using t = 1.10
                 0.7074
```

The t distribution converges on the normal once n exceeds roughly 30 — which is why using z worked at $n = 100$.

In this class you must use the t distribution whenever you substitute the sample variance for the population variance, *even when it barely changes the answer. That is genuinely the distribution of the sampling distribution under those conditions.*

Coming Next

We can now go from one sample to an inferential claim: substitute s^2 for $V[X]$, center a sampling distribution on the sample mean, and use t to set bounds. That is a confidence interval.

The mirror-image question — "given a specific claim about the truth, is my sample consistent with it?" — is a **hypothesis test**, *and that is where we go next.*

Hypothesis Tests

PSCI1801 Chapter 8

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

From Confidence Intervals to Hypothesis Tests

A confidence interval answers: "given my sample, where might the truth plausibly lie?"

$$CI(\alpha) = \left[\bar{X}_n - t_{\alpha/2} \cdot SE,\ \ \bar{X}_n + t_{\alpha/2} \cdot SE\right]$$

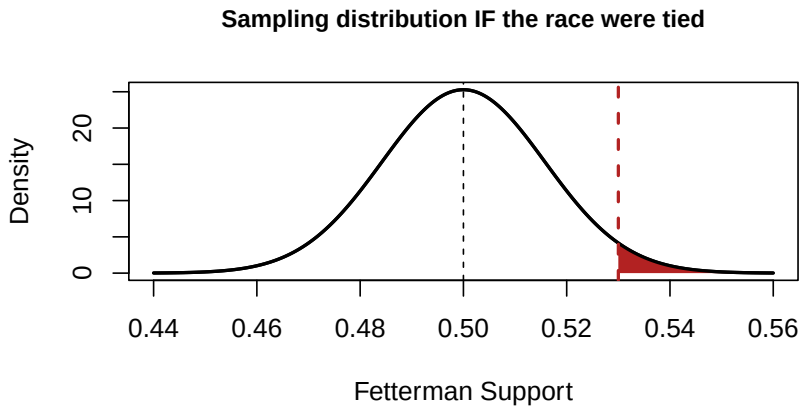
A hypothesis test answers the mirror-image *question: "given a specific claim about the truth, is my sample consistent with it?"*

The Logic: Inference by Contradiction

A poll of 1,000 Pennsylvanians gives Fetterman 53%. Can we rule out a tied race?

We do **not** *know where the sampling distribution is centered. But we* **do** *know how wide it is — the CLT plus the sample variance gives us the standard error.*

So: assume the null is true, draw that sampling distribution, and ask how strange our result looks against it.




```
round(c(sample.mean = xbar,
        p.one.tailed = pnorm(xbar,
                             .5,
                             se,
                             lower.tail = FALSE)), 4)
```

```
 sample.mean p.one.tailed
      0.5300       0.0286
```

Only ~2.9% of samples would be this extreme or more if the race were tied. The data contradict *the null.*

Note the question is never "what is the probability of getting exactly 53%?" — that is zero for a continuous distribution. It is always as extreme or more extreme.

The Five Steps

1. Specify the null and alternative hypotheses.

$$H_0 : P(\text{Fetterman}) = .5$$

$H_a : P(\text{Fetterman}) > .5$ (one-tailed) $H_a : P(\text{Fetterman}) \neq .5$ (two-tailed)

The null can be anything *— a tie, a previous president's approval rating, zero.*

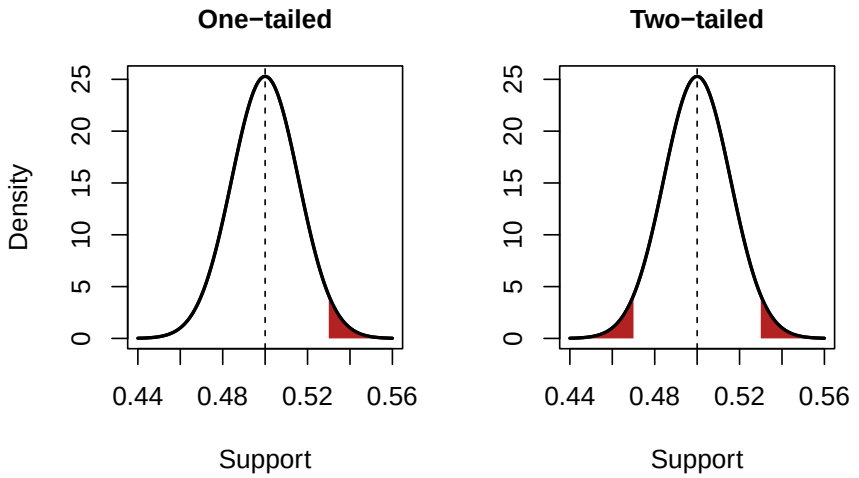
2. Choose a test statistic and α. *α is the* false positive rate *you will accept. Convention is* .05*.*

3. Derive the sampling distribution and center it on the null. *All you need is the standard error. Nothing here is specific to means — this works for any statistic with a known sampling distribution.*

4. Compute the p-value.

5. Reject H_0 if $p \leq \alpha$; otherwise retain it.

One tail or two?



Because the normal is symmetric, the second tail is the same size as the first:

$$p_{\text{two-tailed}} = 2 \times p_{\text{one-tailed}}$$

```
round(c(one.tailed = pnorm(xbar,
                           .5,
                           se,
                           lower.tail = FALSE),
        two.tailed = 2 * pnorm(xbar,
                               .5,
                               se,
                               lower.tail = FALSE)), 4)
```

```
one.tailed two.tailed
    0.0286     0.0573
```

Here the one-tailed test rejects ($.029 < .05$) but the two-tailed test does not ($.057 > .05$). Two-tailed is more conservative, *which is why it is the default in social science.*

Careful with language. *We never "accept" the null. We* fail to reject *it. We are not saying the race* is *tied — only that we lack evidence to overturn that idea.*

A robust way to compute the p-value

Work with the absolute *deviation from the null, converted to standard-error units. That way the formula is correct whether $\bar{X}$ landed above or below H_0:*

$$t = \frac{\bar{X}_n - \mu_0}{s/\sqrt{n}} \qquad p = 2 \times P(T_{n-1} > |t|)$$

```
t.stat <- (xbar - .5) / se
round(c(t = t.stat,
        p = 2 * pt(abs(t.stat),
                   df = 999,
                   lower.tail = FALSE)), 4)
```

```
     t      p
1.9012 0.0576
```

Strictly you should use the t distribution with $n-1$ df. With large n it matches the normal, but use t anyway.

Hypothesis Tests and CIs Are Equivalent

The α in a hypothesis test is the same α as in a confidence interval. If a 95% CI excludes *the null value, a two-tailed test at $\alpha = .05$* rejects *it — always.*

```
library(dplyr)
t.value <- qt(.025, df=999)*-1

reject.null.ci <- rep(NA, 10000)
reject.null.hyp <- rep(NA, 10000)

for(i in 1:10000){
```

```
  samp <- rbinom(1000, 1, .525)
  xbar <- mean(samp)
  se <- sd(samp)/sqrt(1000)

#ci
  ci <- c(xbar-t.value*se, xbar+t.value*se)
  reject.null.ci[i] <- between(.5, ci[1], ci[2])
#hypothesis
  p <- 2*pnorm(abs(xbar - .5)/se, lower.tail=F)
  reject.null.hyp[i] <- p>.05
}

table(reject.null.ci, reject.null.hyp)
```

```
              reject.null.hyp
reject.null.ci FALSE TRUE
         FALSE  3641    0
         TRUE      0 6359
```

Doing This in R

You now know what happens under the hood. In practice, use `t.test()`.

```
t.test(sm.az$approve.biden, mu = .38)
```

```
    One Sample t-test

data:  sm.az$approve.biden
t = 5.1065, df = 3023, p-value = 0.0000003485
alternative hypothesis: true mean is not equal to 0.38
95 percent confidence interval:
 0.4082918 0.4435601
sample estimates:
mean of x
0.4259259
```

Read the output: the `t` *is exactly* $(\bar{X} - \mu_0)/(s/\sqrt{n})$*;* `df` *is* $n - 1$*; the CI shown does not contain* .38*, matching the rejection.*

Defaults to two-tailed; use `alternative = "greater"` *or* `"less"` *for one-tailed.*

Reporting very small p-values: *never round to zero. The curve approaches zero asymptotically but never reaches it. Report "less than 1%."*

The Four Outcomes

	H_0 is true	H_0 is false
Reject H_0	False positive (α)	True positive (power)
Retain H_0	True negative	False negative (β)

α does exactly what you tell it to

Simulate a genuinely tied race and test it 10,000 times:

```
p <- NA
for(i in 1:10000){
samp <- rbinom(1000,1, prob=.5)
#Shortcut with the t.test function:
p[i] <- t.test(samp, mu=.5)$p.value
}

mean(p<.05)
```

```
[1] 0.0529
```

```
mean(p<.1)
```

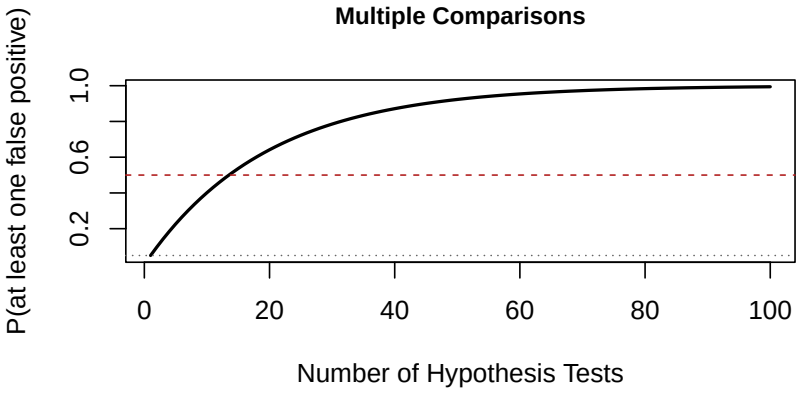
```
[1] 0.092
```

Set a 5% false positive rate, get 5% false positives. That is the deal you signed.

Multiple Comparisons

The 5% applies to one *test. Run many, and the chance that* at least one *is a false positive climbs fast:*

$$P(\text{at least one false positive}) = 1 - (1 - \alpha)^m$$



Ten tests → ~40% chance of a false positive. Eighty tests → near certainty.

This is a major driver of the **replication crisis**. *In one large psychology replication effort, of 97 original studies with significant effects, only 36% replicated, with effect sizes averaging* half *the original magnitude.*

Two versions of how this happens:

- **p-hacking** *(uncharitable): deliberately mining data for the comparison that turns up significant.*
- **The "garden of forking paths"** *(Gelman, more charitable): researchers make defensible choices at each step — which subgroup, which cutpoint, which scale — but because those choices respond to the data, and because they stop when they find significance, the result is the same.*

Two responses

Bonferroni correction. *Test against α/m instead of α, where m is the number of comparisons.*

```r
m <- 10
coin <- c(0,1)
any.sig <- NA

for(j in 1:1000){
  sig <- NA
  for(i in 1:m){
    flips <- sample(coin, 100, replace=T)
    #swap .05 for (.05/m) for the Bonferroni correction
    sig[i] <- t.test(flips, mu=.5)$p.value <.05
  }
  any.sig[j] <- any(sig)
}
mean(any.sig)
```

```
[1] 0.446
```

Correction pulls the family-wise error rate back to about 5%. But it only works when you can actually count *your comparisons — often you cannot.*

Pre-registration. *Plan the analysis and hypotheses* before *collecting data. You are locked in, choices cannot respond to results, and the ordinary rules apply. This is now the dominant norm in much of science.*

False Negatives

The other error: failing to reject when we should. Kelly's true support was 52.4%. Simulate polls of $n = 2613$ *and test* $H_0 : p = .5$*:*

```r
set.seed(19104)
false.negative <- rep(NA,10000)

n <- 2613
alpha <- .05

for(i in 1:10000){
samp <- rbinom(n,1, .524)
s <- sd(samp)
xbar <- mean(samp)
se <- s/sqrt(n)
t.val <- (xbar-.5)/se
```

```
p <- 2*pt(abs(t.val), df=n-1, lower.tail=F)
false.negative[i] <- p>=alpha
}
mean(false.negative)
```

```
[1] 0.3215
```

About 32% false negatives — far higher than our carefully chosen 5% false positive *rate. We call this β.*

A false negative is usually less damaging than a false positive (better to wrongly say a drug doesn't work than wrongly say it does), but it is often a far larger *problem in practice.*

Coming Next

We have the full machinery of a hypothesis test. Next: **power** *— the flip side of β, telling us when a study can even detect a real effect — and the* **bootstrap**, *a computational way to build a sampling distribution around* any *statistic, not just the mean.*

Power & the Bootstrap

PSCI1801 Chapter 9

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Power

Four possible outcomes of a hypothesis test:

	H_0 is true	H_0 is false
Reject H_0	False positive (α)	True positive (**power**)
Retain H_0	True negative	False negative (β)

α is the only *one we set by hand. Convention is 5% — which is actually very high. On the NBC News decision desk we reject the null that a race is tied. In 2024 we called ~600 races; at $\alpha = .05$ we would get $600 \times .05 = 30$ races wrong. We would lose our jobs. Even our actual $\alpha = .005$ implies 3 wrong calls.*

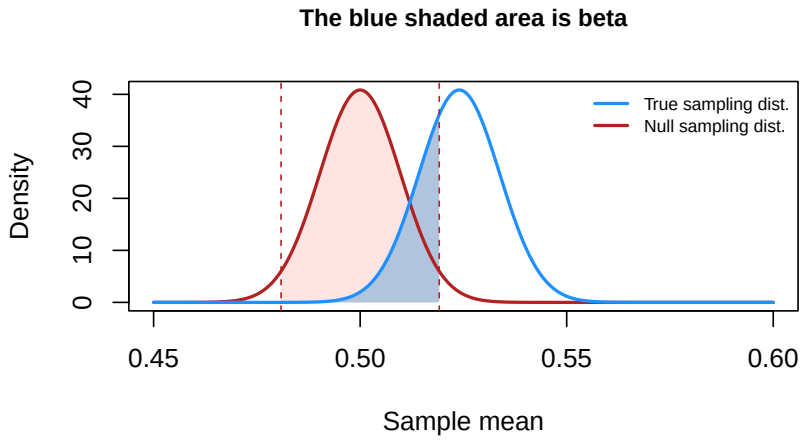
β is not something we set — but we can derive it exactly.

What Is β?

Suppose Kelly's true support is 52.4% and we poll $n = 2613$. Two distributions matter:

- *The* **true** *sampling distribution (blue), centered on .524 — this is what actually produces our polls.*
- *The* **null** *sampling distribution (red), centered on .50 — this is what sets the rejection cutoffs.*

The dashed lines are the α cutoffs, computed from the null. *Between them is the region where we* fail to reject. *Ask:* how much of the TRUE (blue) curve falls in that retain region? *That area is β.*




```
n <- 2613; mu <- .524; null <- .5; alpha <- .05
se <- sqrt(mu * (1 - mu) / n)
# Rejection bounds come from the NULL distribution...
lo <- qnorm(alpha/2, null, se)
hi <- qnorm(alpha/2, null, se, lower.tail = FALSE)
# ...but the probability comes from the TRUE distribution.
round(pnorm(hi, mu, se) - pnorm(lo, mu, se), 4)
```

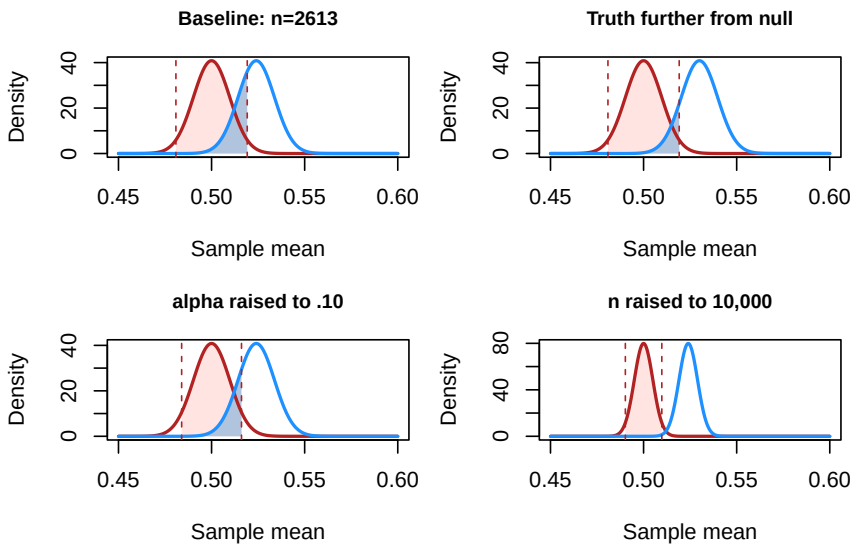
```
[1] 0.3098
```

About 31% — matching the simulation from last chapter. **31% of the polls produced by the truth will fail to reject a tied race, even though Kelly really is ahead.** *Data like this are* **under-powered**.

$$\text{power} = 1 - \beta$$

Power is the probability of rejecting the null when you should — *your ability to detect something that is really there.*

Four Levers on β




```
round(c(baseline = b1,
        bigger.effect = b2,
        higher.alpha = b3,
        bigger.n = b4), 3)
```

```
    baseline bigger.effect  higher.alpha      bigger.n
       0.310         0.133         0.208         0.002
```

1. A bigger true effect. *The further the truth sits from the null, the less overlap. Real, but useless — we do not control the truth.*

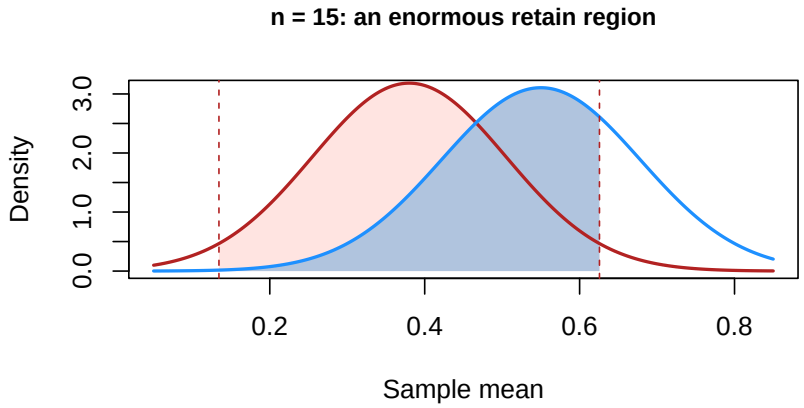
2. A higher α. *Accept more false positives, get fewer false negatives.* **Do not make this trade.** *Better to say a drug doesn't work when it does than to say it works when it doesn't.*

3. A larger n. *The workhorse. Since $SE = s/\sqrt{n}$, more data narrows* both *curves, shrinking their overlap.*

4. A smaller s. *Harder to control, but real. Feeling thermometers (rate this politician 0–100) are answered terribly — huge random variation. All else equal, a noisy measure means higher β.*

Underpowered and Overpowered

Biden's peak approval was 55%. A firm surveys 15 people *and reports it cannot tell whether he beats Trump's exit approval of 38%.*




```
round(pnorm(bounds[2], hA, seA) -
      pnorm(bounds[1], hA, seA), 3)
```

```
[1] 0.721
```

With 15 people, the band needed to hold the false positive rate at 5% is enormous. It is more likely than not *that we get a false negative.*

The fallacy to watch for: *"we ran an experiment and found nothing, therefore nothing is there." A power analysis often reveals a 20–30% expected false negative rate.* You cannot claim nothing is happening when your method returns "nothing" 30% of the time.

Can you have too much power?

Not exactly — but it misleads. True support 38.5%, null 38%, and a poll of one million:

```
n <- 1e6; hA <- .385; h0 <- .38
seA <- sqrt(hA*(1-hA)/n); se0 <- sqrt(h0*(1-h0)/n)
bounds <- c(h0 + qnorm(.025)*se0, h0 - qnorm(.025)*se0)
round(pnorm(bounds[2], hA, seA) -
      pnorm(bounds[1], hA, seA), 6)
```

```
[1] 0
```

Essentially zero chance of a false negative — so a half-point difference is "statistically significant."

Statistical significance says nothing about magnitude or importance. *Nobody thinks 38.5% approval is substantively different from 38%.*

The red meat disaster

The WHO classified processed meat as a Group 1 carcinogen, *alongside asbestos and tobacco. Group 1 is not a measure of how* severe *a carcinogen is — it is a statement of* **statistical significance**. *Enough studies had accumulated to reject the null of no effect.*

It got worse in the reporting. "Eating 50g of processed meat a day raises colon cancer risk 18%" is a relative risk. *A 3% baseline becomes* $3 \times 1.18 = 3.54\%$ *— not* $3 + 18 = 21\%$.

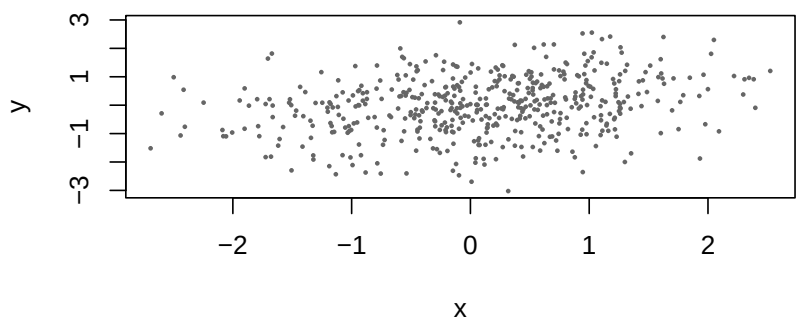
For comparison: a lifelong non-smoker has ~0.5% lung cancer risk; a lifelong smoker ~15%. That is roughly 30 times *the risk — a relative increase near 3000%, against meat's 18%.*

If everyone misinterprets what you said as a scientist, that's your fault, not theirs.

The Bootstrap

The CLT handed us a formula, $s/\sqrt{n}$, *plus a guarantee of normality. That covers the mean. What about a statistic with* no formula at all?

The motivating problem



```r
correlation <- cor(d$x, d$y)
round(correlation, 4)
```

```
[1] 0.257
```

```r
names(correlation)  # is there a standard error in there?
```

```
NULL
```

`NULL`. No standard error. The textbook has no formula for one either.

(A note you will need repeatedly: `cor()` returns `NA` if either vector has missing values, and `na.rm = TRUE` is not *the fix. Use `use = "pairwise.complete"`.)*

The key insight

> **Because our sample is an iid sample of the population, a random sample *of our sample* is a new sample.**

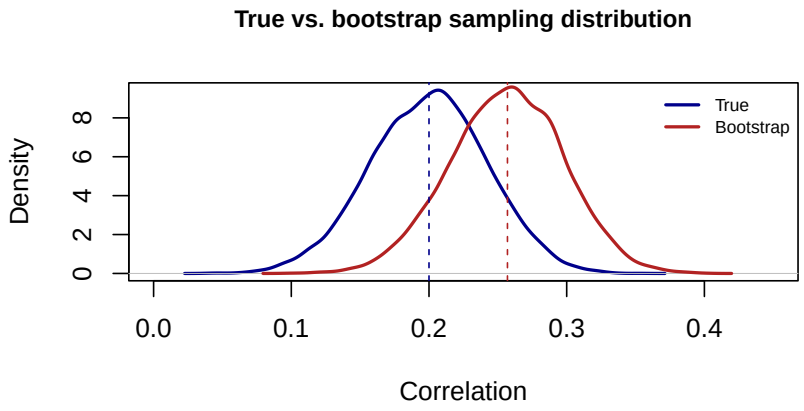
The range of possibilities in the population is already latent in our data. So resample with replacement *to generate new datasets and recompute the statistic in each.*

Why with replacement? Sampling without *replacement just reorders the rows — you get the identical dataset back and the identical correlation. Repeated entries are what create the variation.*

```r
set.seed(19104)
bs.samp.dist <- rep(NA, 10000)
for(i in 1:10000){
  bs.samp <- d[sample(1:nrow(d), replace=T),]
  bs.samp.dist[i] <- cor(bs.samp$x, bs.samp$y)
}
sd(bs.samp.dist)
```

```
[1] 0.04115306
```

Does that match the truth? Here we generated the data, so we can build the real sampling distribution by resampling the population:




```
sd(cor.samp.dist)
```

```
[1] 0.04230724
```

```
sd(bs.samp.dist)
```

```
[1] 0.04115306
```

They are not centered in the same place — and that is fine. *The true distribution centers on the population value (.2); the bootstrap centers on* our sample's *value. We do not care where a sampling distribution sits. We care about its* width *— and the standard errors are identical.*

What you do with it

```
#percentile CI
round(quantile(bs.samp.dist, c(.025, .975)), 4)
```

```
  2.5%  97.5%
0.1744 0.3343
```

```r
z <- correlation / sd(bs.samp.dist) # or a hypothesis test
round(2 * pnorm(z, lower.tail = FALSE), 5)
```

```
[1] 0
```

Four Ways to Get a Sampling Distribution

1. *Know the population, compute a standard error from its features.* **(Impossible in the real world.)**
2. *Know the population, repeatedly sample and compute the statistic.* **(Impossible in the real world.)**
3. *Use your sample plus a formula someone has worked out.* **(Requires a formula to exist.)**
4. *Generate bootstrap samples and compute the statistic in each.* **(Always works.)**

In practice only (3) and (4) are available. If (3) exists, always prefer it. *Otherwise (4) is there.*

A Real Finding

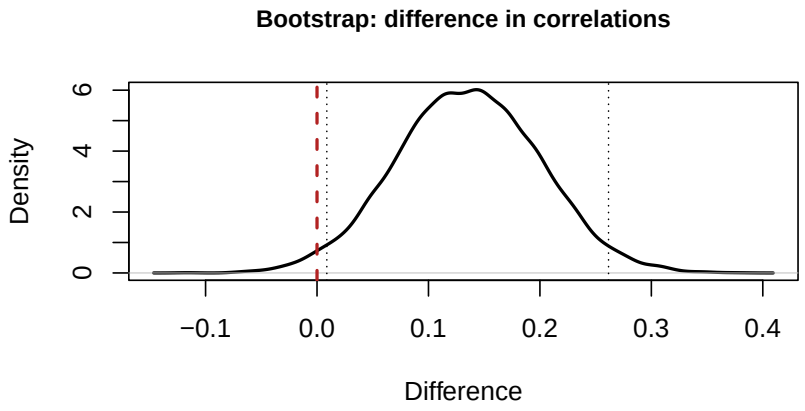
Does ideological self-placement predict voting for a centrist in the 2020 Democratic primary? Following Jefferson (2024), the cultural meaning of "liberal" and "conservative" may differ enough for Black Americans that they are effectively answering a different question.

```r
w <- anes$white.v.black == 1
b <- anes$white.v.black == 0
cw <- cor(anes$ideology[w],
          anes$vote.centrist.dem[w],
          use = "pairwise.complete")
cb <- cor(anes$ideology[b],
          anes$vote.centrist.dem[b],
          use = "pairwise.complete")
round(c(white = cw, black = cb, difference = cw - cb), 4)
```

```
     white      black difference
    0.2818     0.1446     0.1372
```

Substantially lower among Black respondents. But the ANES has only ~700 Black respondents, not all of whom voted in the primary — so how much would that difference bounce around across samples?

Check the four options: we don't know the population (rules out 1 and 2). Is there a formula for the standard error of a difference between two correlations? *There isn't even one for a* single *correlation. So: bootstrap.*




```
#CI excludes zero
round(quantile(bs.delta, c(.025, .975)), 4)
```

```
  2.5%  97.5%
0.0088 0.2616
```

```
round(2 * pnorm(cor.delta / sd(bs.delta),
                lower.tail = FALSE), 4)
```

```
[1] 0.0327
```

The interval excludes zero, and there is about a 3% chance of a difference this extreme if the two correlations were truly equal.

That is a genuinely new finding, produced with a statistic that has no standard error formula at all.

Coming Next

Power tells us what a test can and cannot detect at a given n. The bootstrap builds a sampling distribution around any *statistic. Both generalize the logic we built for the sample mean. Next we develop* **covariance and correlation** *properly — the statistics used as examples here — and then* **regression**.

Covariance and Correlation

PSCI1801 Chapter 10

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Moving to Relationships

So far only one estimator: the sample mean. But we usually care about how variables move together *— how they* **co**-vary. *Three tools, in order of power: difference in means, covariance, correlation.*

Difference in Means

An experiment: $\boldsymbol{x}$ *is treatment (1) or control (0),* $\boldsymbol{y}$ *is the outcome. Because treatment is* randomly assigned, *any difference in* $\boldsymbol{y}$ *must be due to treatment.*

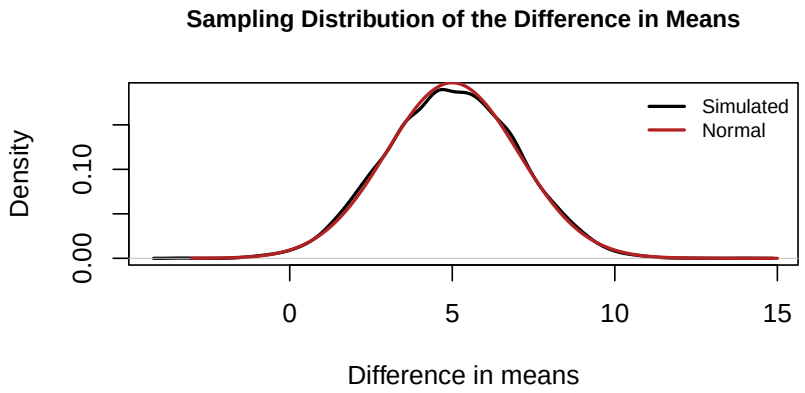
$$H_0 : \bar{X}_1 - \bar{X}_0 = 0 \qquad H_a : \bar{X}_1 - \bar{X}_0 \neq 0$$

Rearranging into a single number *is the trick — now we have one test statistic to work with.*

```
diff.means <- mean(dat$y[dat$x == 1]) -
              mean(dat$y[dat$x == 0])
round(diff.means, 3)
```

```
[1] 5.489
```

Every new sample gives a different difference. What is the sampling distribution?



Normal, centered on the true difference (5), with a standard deviation of about 2 — which is, as always, the **standard error**.

The formula generalizes the one-mean case. Now there are two *populations and* two *group sizes:*

$$SE = \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \qquad \longrightarrow \qquad \widehat{SE} = \sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}$$

And, as before, substituting sample variances makes it t distributed:

$$\frac{(\bar{X}_1 - \bar{X}_0) - (\mu_1 - \mu_0)}{\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}} \sim t_v$$

The degrees of freedom v has a genuinely horrible formula. You will never be asked to compute it — R does it.

```
t.test(dat$y ~ dat$x)
```

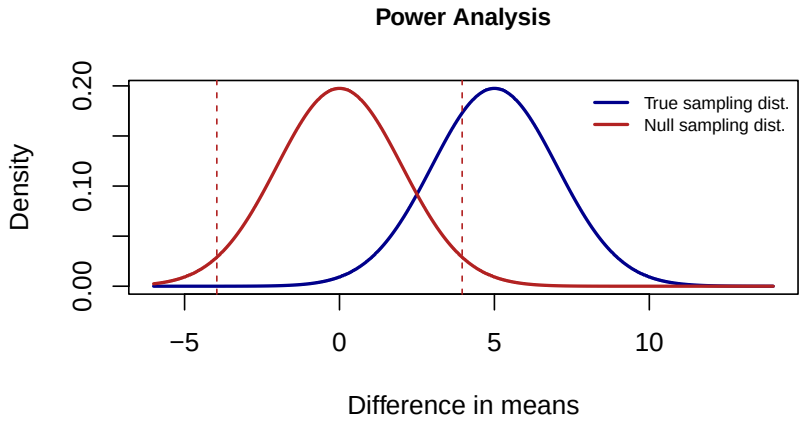
```
    Welch Two Sample t-test

data:  dat$y by dat$x
t = -3.1014, df = 77.848, p-value = 0.002683
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -9.012638 -1.965384
sample estimates:
mean in group 0 mean in group 1
       22.44682        27.93583
```

The big lesson: *deriving the standard error was harder, but once you have* a test statistic, a standard error, and a distribution shape, *the hypothesis test is* identical *to the one-mean case. That is true for every estimator in this course.*

A difference in means is really just the covariance between a binary x *and a continuous* y *— which is how it connects to everything below.*

Power for a Difference in Means

Two distributions matter: the one centered on the null *(which sets the rejection cutoffs) and the one centered on the* truth *(which actually generates our data).*



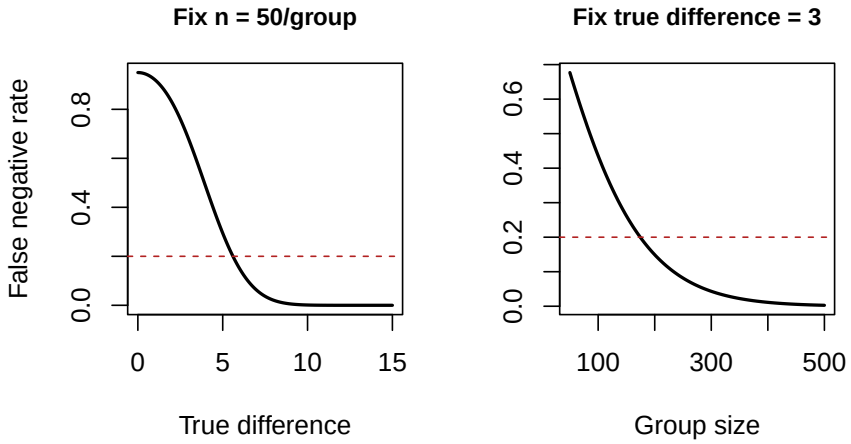
The false negative rate β *is the area under the* blue *curve that falls* between *the red cutoffs:*

```
round(pnorm(1.96*2.02, 5, 2.02) -
      pnorm(-1.96*2.02, 5, 2.02), 3)
```

```
[1] 0.303
```

A 30% false negative rate with $n = 100$*. Those are terrible odds — you would waste the study.* **This is why you run a power analysis before collecting data.**

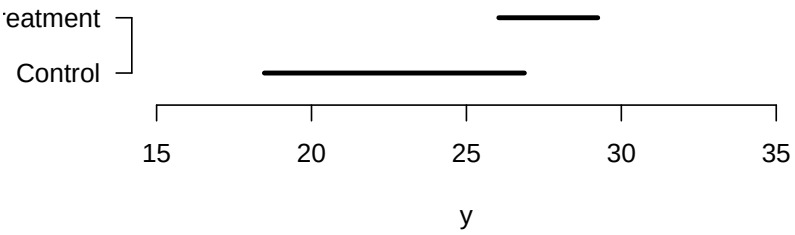
You need two ingredients: a guess at the true effect *and* a guess at the standard error *(which needs* n *and the variance — look up the variance from prior studies, or assume a pessimistic value).*



Convention is 20% false negatives (80% power). Where that line crosses gives the **minimum detectable effect size** *(left) or the required n (right).*

Overlapping Confidence Intervals Are Not a Test

A trap you will see made constantly. Two group means with individual 95% CIs that overlap:




```
round(t.test(ds$y ~ ds$x)$p.value, 4)
```

```
[1] 0.0298
```

The intervals overlap, yet the difference is significant. Simulating 10,000 times:

	t-test rejects	t-test retains
CIs don't overlap	4,340	**0**
CIs overlap	2,509	3,151

The rules: *if two 95% CIs do* **not** *overlap, the test* will *reject (that top-right cell is zero). If they* **do** *overlap, you learn* nothing — *over a third of real differences hide there. Best not to use the eyeball method at all.*

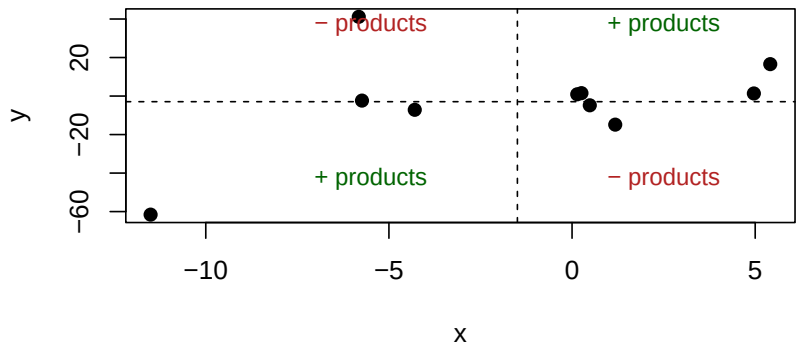
Covariance

With two continuous *variables there are no groups to compare. We need a summary of how they move together.*

$$\hat{\sigma}_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Take each point's deviation from $\bar{x}$, times its deviation from $\bar{y}$, and average the products. **Nothing is squared** — *so unlike variance, covariance can be negative.*

The intuition is quadrants, split at the two means:



- *Points in the* **lower-left and upper-right** *make* positive *products (both deviations same sign).*
- *Points in the* **upper-left and lower-right** *make* negative *products.*
- *The deeper into a quadrant, the bigger the contribution.*

The single covariance number summarizes the relative frequency and size of products across the four quadrants.

Correlation

Covariance has a fatal flaw: it depends on scale.

```
cov(a$x, a$y)
```

```
[1] 60.72786
```

```
cov(a$x*100, a$y*100)
```

```
[1] 607278.6
```

The relationship did not get 10,000 times stronger — the units changed. Fix it by standardizing:

$$\hat{\rho}_{x,y} = \frac{\hat{\sigma}_{x,y}}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}}$$

```
cor(a$x, a$y)
```

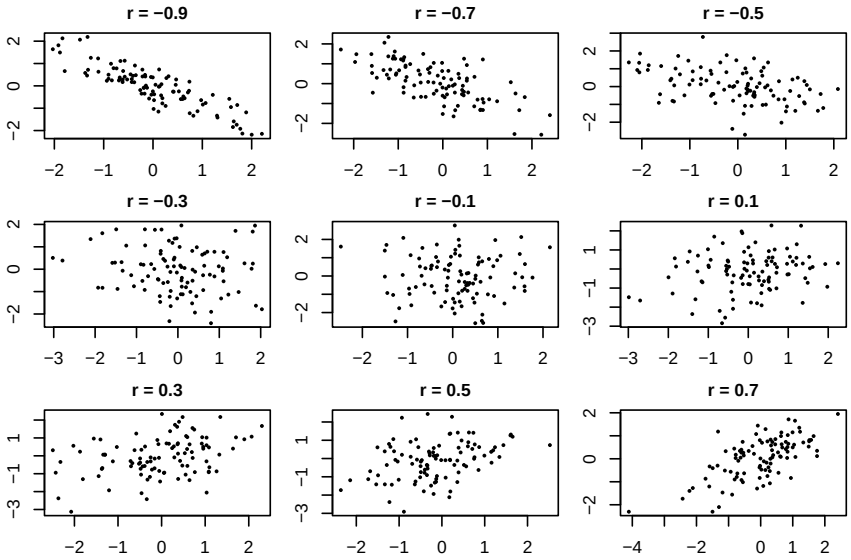
```
[1] 0.4461244
```

```
cor(a$x*100, a$y*100)
```

```
[1] 0.4461244
```

Identical. Correlation is bounded: $-1 \leq \rho \leq 1$.

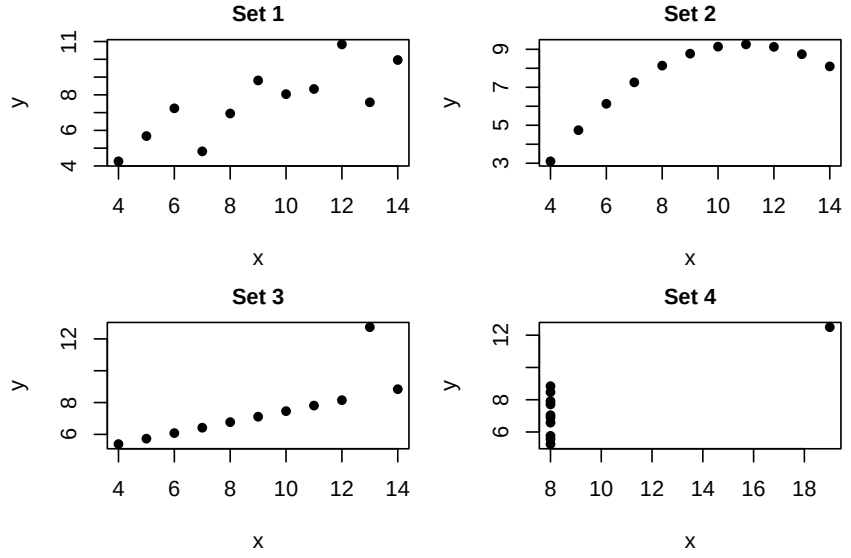
Why does perfect correlation give exactly 1? If every point has $(x_i - \bar{x}) = (y_i - \bar{y})$*, then the covariance formula and the variance formula are* the same equation *— so the ratio is 1.*



Correlations are genuinely hard to eyeball — try guessthecorrelation.com.

Anscombe's quartet

Four datasets. All four have a correlation of about **0.82**.




```
cor(anscombe$x1, anscombe$y1)
```

```
[1] 0.8164205
```

```
cor(anscombe$x2, anscombe$y2)
```

```
[1] 0.8162365
```

```
cor(anscombe$x3, anscombe$y3)
```

```
[1] 0.8162867
```

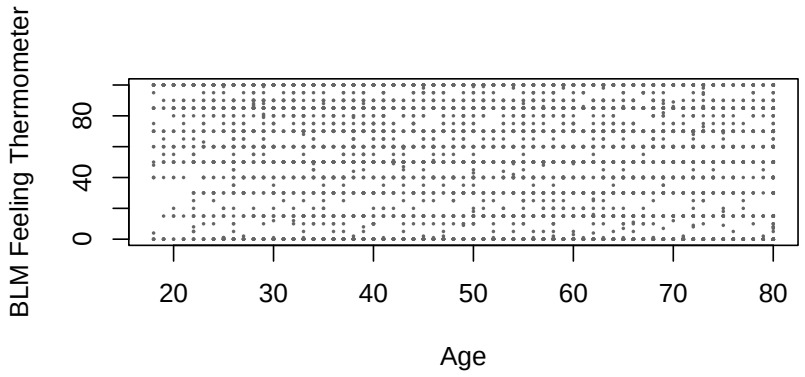
```
cor(anscombe$x4, anscombe$y4)
```

```
[1] 0.8165214
```

Only Set 1 is the loose cloud you pictured. Set 2 is a perfect curve *(a straight-line summary is the wrong tool). Set 3 is a perfect line with one outlier dragging the coefficient down. Set 4 is a vertical stack where a* single point *invents the relationship.*

Always plot your data before trusting one number to describe it.

Real Data: Age and BLM Support



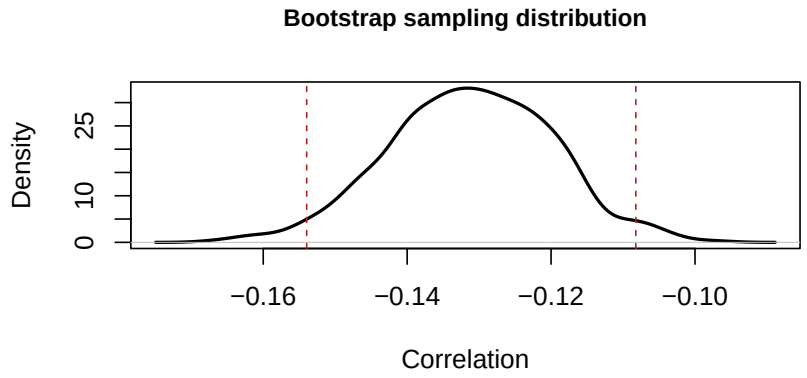
A blob — 7,000 people over-plotted on a 0–100 scale. This is exactly where one summary number helps.

```
round(cor(anes$age,
          anes$blm.therm, use = "pairwise.complete"), 4)
```

```
[1] -0.1307
```

A slight negative relationship: older respondents are, on average, a bit cooler toward BLM. Not determinative.

Can we trust it? There is **no formula** *for the standard error of a correlation. A sampling distribution certainly exists — it must — but we have no math for it. So we* **bootstrap**: *resample the data with replacement, recompute the correlation each time.*




```
round(quantile(cor.bs, c(.025, .975)), 4)
```

```
   2.5%   97.5%
-0.1540 -0.1082
```

The 95% interval does not contain 0, so we are confident the true correlation is not zero.

Coming Next

Covariance and correlation summarize *how two variables move together — but summarizing is not* modeling. **Regression** *builds on correlation to give effect sizes in the actual units of our data, the ability to hold other variables constant, and predictions.*

Regression

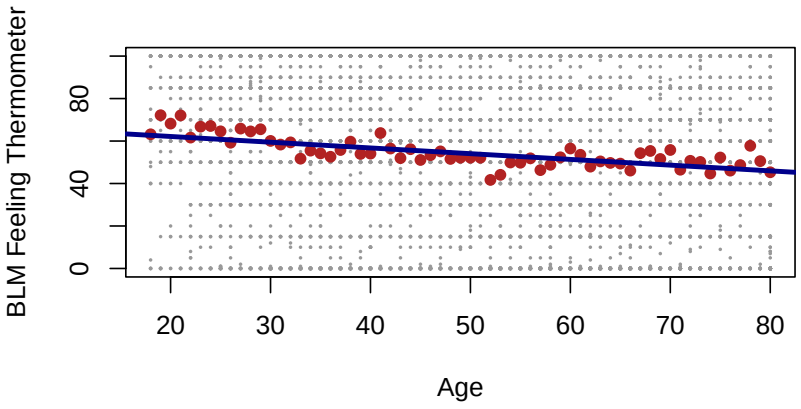
PSCI1801 Chapter 11

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

From Correlation to a Line

Last chapter we summarized age vs. BLM feeling thermometer with a correlation of -0.13. *Now we want something more concrete: a* line.

One naive option — compute the mean of $\boldsymbol{y}$ *at every value of* $\boldsymbol{x}$ *and connect the dots:*



The red line follows the data religiously — but we would never claim that every wiggle exists in the population. *That is* **over-fitting**: *treating the quirks of one random sample as truth. We want to* smooth *out sampling noise into a single summary.*

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_i$$

How Does OLS Choose the Line?

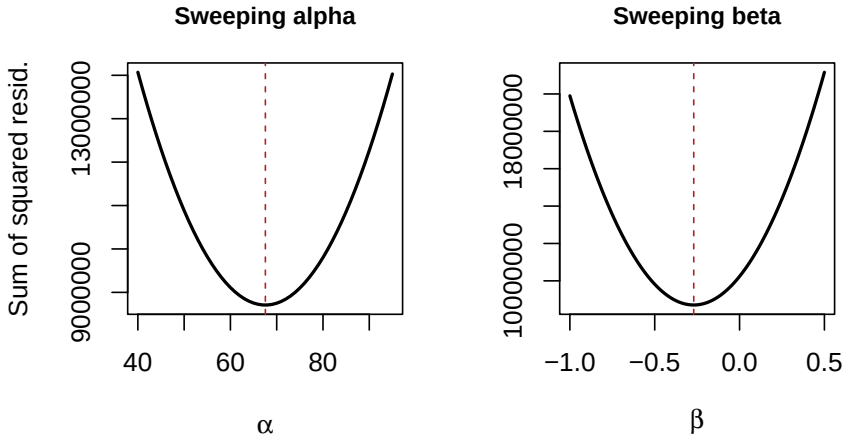
For each point, the **residual** *is the vertical distance from the point to the line:*

$$\hat{u}_i = y_i - \hat{y}_i$$

Ordinary Least Squares *chooses* $\hat{\alpha}$ *and* $\hat{\beta}$ *to* minimize the sum of squared residuals:

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2$$

Proof by brute force: hold one parameter at its OLS value and sweep the other.



Both curves bottom out exactly at the OLS values. Calculus does the same thing: take the partial derivative with respect to each parameter, set it to zero (the slope of the curve is flat at a minimum), and solve:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Covariance, correlation, and regression are all the same thing

Look hard at that $\hat{\beta}$*. Divide numerator and denominator by* $n - 1$ *and it becomes:*

$$\hat{\beta} = \frac{\hat{\sigma}_{x,y}}{\hat{\sigma}_x^2} \qquad \text{compare} \qquad \hat{\rho}_{x,y} = \frac{\hat{\sigma}_{x,y}}{\sqrt{\hat{\sigma}_x^2 \hat{\sigma}_y^2}}$$

A regression coefficient is the covariance divided by the variance of x. A correlation is the covariance divided by both standard deviations. *They are different* scalings *of the same quantity — which is why they can never disagree about the* direction *of a relationship.*

```
lm(anes$blm.therm ~ anes$age)
```

```
Call:
lm(formula = anes$blm.therm ~ anes$age)

Coefficients:
(Intercept)     anes$age
      67.55        -0.27
```

```
cov(anes$blm.therm, anes$age)/var(anes$age)
```

```
[1] -0.2699798
```

Interpreting the Output

```
m <- lm(blm.therm ~ age, data = anes)
ctab(m)
```

```
             Est    SE      t      p
(Intercept) 67.55 1.329  50.84 <0.001
age         -0.27 0.024 -11.08 <0.001
```

The intercept *is the predicted y when $x = 0$ — here, a newborn's BLM thermometer. Mechanically necessary (you cannot draw a line without it), substantively meaningless. Its hypothesis test is equally meaningless.*

The slope *is rise over run: a* one unit *increase in x changes y by $\hat{\beta}$, on average. Here each additional year of age is associated with a* 0.27 *point* drop.

"One unit" always means one unit in the number system — *1 to 2, 100 to 101. Internalize this now; it never changes. You can scale it freely: 10 years → 2.7 points.*

Rescaling x changes beta but not the relationship

```
anes$age.months <- anes$age*12
m2 <- lm(blm.therm ~ age.months, data=anes)
coef(m2)["age.months"]
```

```
 age.months
-0.02249832
```

```
coef(m2)["age.months"]*12
```

```
age.months
-0.2699798
```

The relationship did not weaken — we are now measuring the effect of one month. *This follows directly from $\hat{\beta} = \hat{\sigma}_{x,y}/\hat{\sigma}^2_x$: regression standardizes by whatever scale x happens to be on.*

Regression With a Binary Independent Variable

Everything OLS needs is numeric *variables. Let* `liberal` *be 1 for liberals, 0 otherwise.*

```
anes$liberal <- ifelse(anes$ideology < 4, 1, 0)
m3 <- lm(blm.therm ~ liberal, data = anes)
round(coef(m3), 2)
```

```
(Intercept)     liberal
      38.05       40.38
```

- *Intercept = average BLM thermometer among* non-liberals *(where $x = 0$).*

- *Slope = the effect of a one unit change, i.e. moving from one group to the other.*
- *Predicted value for liberals:* $38.05 + 40.38(1) = 78.43$.

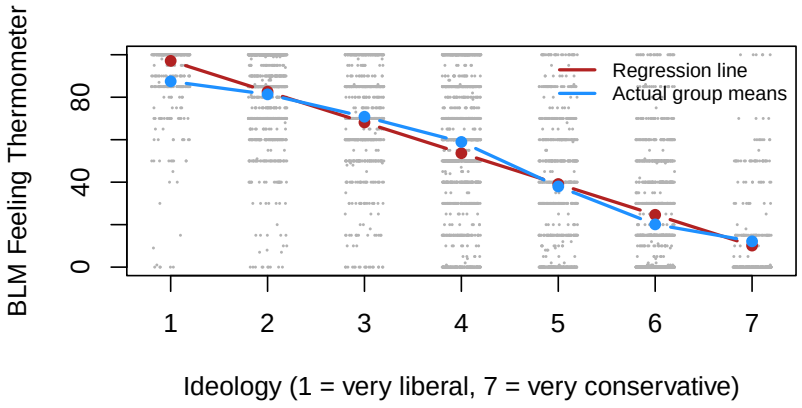
```
round(t.test(anes$blm.therm ~ anes$liberal)$estimate, 2)
```

```
mean in group 0 mean in group 1
          38.05           78.44
```

OLS with a binary x exactly recovers the two group means, with $\hat{\beta}$ as the difference between them — *it* is *a difference in means test. This is why indicator variables work so beautifully with regression: a one-unit shift is exactly group membership.*

Ordinal Variables: The Straight-Line Assumption

Ideology runs 1 (very liberal) to 7 (very conservative). Regression assumes every variable is continuous and evenly spaced.

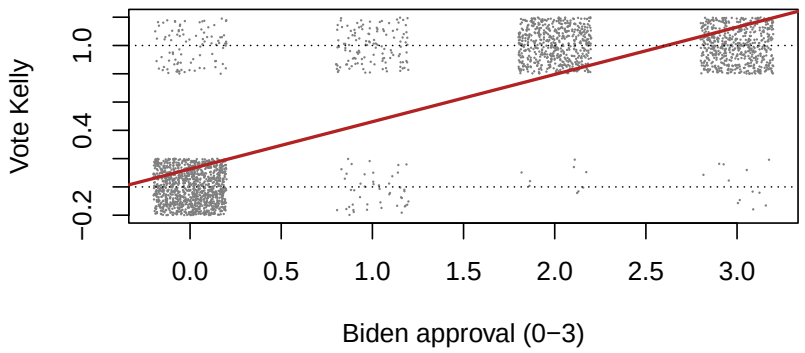


They are not *the same. The regression line is forced straight; the group means are not. Regression treats the jump from 1$→2asidenticalto3→$4, and the data say otherwise (the middle jumps are bigger).*

Neither is right or wrong — but you must know what assumption you are making.

Regression With a Binary Dependent Variable

Put a 0/1 variable on the left *side and OLS becomes a* **Linear Probability Model (LPM)**. *Because the mean of a binary variable is a probability, every coefficient is now read in terms of probability.*




```
round(coef(m5), 3)
```

```
(Intercept)   biden.num
      0.127       0.334
```

```
#predicted probabilities
round(coef(m5)[1] + coef(m5)[2] * 0:3, 3)
```

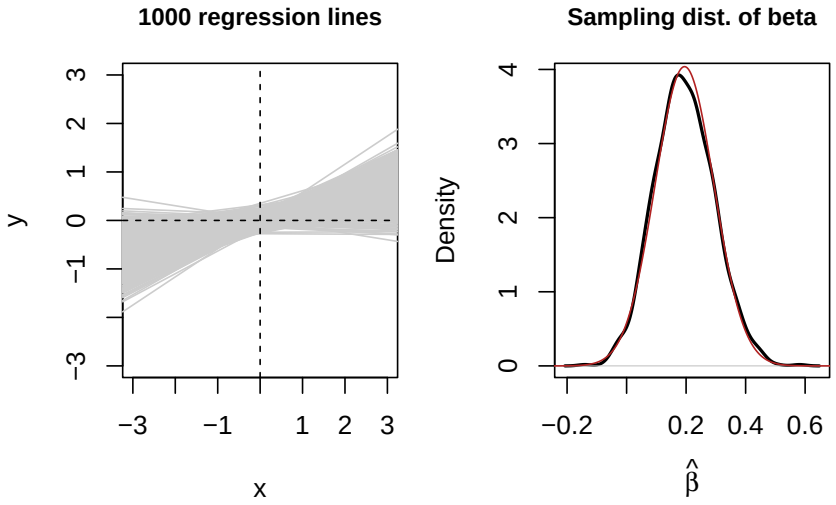
```
[1] 0.127 0.461 0.795 1.130
```

Each step up the approval scale raises the probability of voting Kelly by 33 points. But look at the last predicted value: **112%**. *OLS does not know or care that probabilities live in* $[0, 1]$.

This matters a lot for **prediction** *(e.g. a likely-voter model must stay in* $[0, 1]$ *— use logit or probit, beyond this course) and much less for* **explanation** *(here I mainly want to say Biden approval has a strong positive effect).*

The Standard Error of a Regression Coefficient

A $\hat{\beta}$ is an estimate like any other — it has a sampling distribution. Simulate 1,000 samples from a population where we control the covariance:



Two things. First, the sampling distribution is normal, *centered on the true slope. Second, the lines form a* **"double fan"** *converging at the middle.*

That is no accident — **every regression line passes through** $(\bar{x}, \bar{y})$:

$$\hat{y} = \hat{\alpha} + \hat{\beta}\bar{x} = (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}\bar{x} = \bar{y}$$

So we are most certain about the line near the means, and least certain out at the extremes. This will matter enormously for prediction.

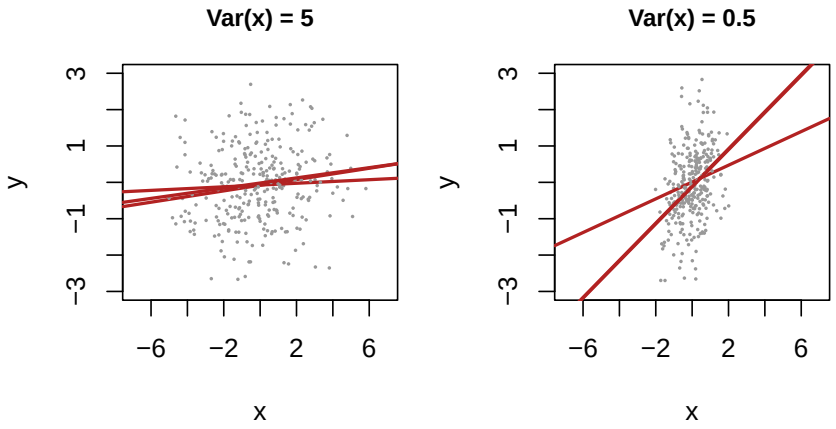
What drives the standard error

$$SE_{\hat{\beta}} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Average squared residual on top, variation in x on the bottom. So:

- *More $n \rightarrow$* **smaller** *SE.*
- *Smaller residuals (tighter relationship) $\rightarrow$* **smaller** *SE.*
- *More variance in $x \rightarrow$* **smaller** *SE.*

That last one surprises people. More spread-out x values give more **leverage** *— when x is bunched up, small random wiggles in y swing the line wildly.*



Left: wide x, the three lines nearly coincide. Right: narrow x, they scatter.

Then it is all the same as before

Degrees of freedom for a regression are n minus the number of coefficients estimated (including the intercept). With $n = 100$ and two coefficients, $df = 98$.

$$t = \frac{\hat{\beta} - 0}{SE_{\hat{\beta}}} \sim t_{n-k}$$

Once you have an estimate and a standard error, everything we have done all semester applies unchanged *— hypothesis tests, confidence intervals, power. You are cooking with gas.*

Coming Next

You now know where $\hat{\alpha}$ and $\hat{\beta}$ come from, how to interpret them, and how their hypothesis tests work. The defining feature of OLS, though, is that it can handle many *variables at once — letting us hold other factors constant. That is next.*

Introduction to Bayesian Statistics

PSCI1801 Chapter 13

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

The Fundamental Difference

Everything this semester has been **frequentist** *statistics. The sampling distribution describes all possible test statistics across hypothetical repeated samples, and we ask:*

> **Frequentist:** *If the null hypothesis is true, how likely are these data?*

Bayesian statistics flips the conditioning entirely:

> **Bayesian:** *Given the data I observed, how likely is the hypothesis?*

These are not the same question. *A p-value of 0.03 does* not *mean a 3% chance the null is true. It means a 3% chance of data this extreme* if the null were true.

The Bayesian framework gives you what most people think *p-values give them — a direct probability statement about a hypothesis: "there is an 85% probability the effect is positive." You* cannot *say that in the frequentist framework, though people do it constantly without realizing they are breaking the rules.*

Deriving Bayes' Theorem

We already have every tool needed. Start with conditional probability, from the probability chapter:

$$P(A \mid B) = \frac{P(A \,\&\, B)}{P(B)} \qquad P(B \mid A) = \frac{P(B \,\&\, A)}{P(A)}$$

Rearranging the second gives $P(A\&B) = P(B \mid A)P(A)$. *Substitute into the first:*

$$\boxed{P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}}$$

The medical test

You test positive. What is the probability you have the disease?

$$P(\text{Dis} \mid \text{Pos}) = \frac{P(\text{Pos} \mid \text{Dis})\, P(\text{Dis})}{P(\text{Pos})}$$

- $P(\text{Pos} \mid \text{Dis}) = .95$ — *the true positive rate.*
- $P(\text{Dis}) = .01$ — *disease prevalence in the population.*
- $P(\text{Pos})$ — *trickier: there are* two routes *to a positive test.*

$$P(\text{Pos}) = \underbrace{(.01)(.95)}_{\text{true positive}} + \underbrace{(.99)(.04)}_{\text{false positive}} = .0095 + .0396 = .0491$$

$$P(\text{Dis} \mid \text{Pos}) = \frac{.95 \times .01}{.0491} = .1935$$

Only a **19%** *chance you have the disease. Surprisingly low — because the false positive route (*.0396*) dwarfs the true positive route (*.0095*) when the disease is rare.*

Bayes is iterative. *Test positive a second time? Your* prior *is no longer 1% — it is now 19.35%. You have effectively entered a new population.*

$$P(\text{Dis} \mid \text{Pos}_2) = \frac{.95 \times .1935}{(.1935)(.95) + (.8065)(.04)} = \frac{.1838}{.2161} = .851$$

You always need at least two hypotheses

What is $1 - .1935 = .8065$*? It is* $P(\textit{NoDis} \mid \textit{Pos})$ — *run through Bayes with the same denominator:*

$$P(\text{NoDis} \mid \text{Pos}) = \frac{(.04)(.99)}{.0491} = .8065$$

Two things follow, and the second is the key to everything below:

1. *You cannot test a* single *hypothesis. You need a set that sums to 1.*
2. **The denominator is identical across hypotheses — it is just the sum of all the numerators.**

Vocabulary

Term	Symbol	In the disease example
Posterior	$P(H \mid D)$	P(disease given positive test)
Likelihood	$P(D \mid H)$	P(positive test given disease)
Prior	$P(H)$	Disease prevalence
Marginal likelihood	$P(D)$	Overall P(positive test)

The likelihood *is what frequentist statistics computes. Since every hypothesis is divided by the same $P(D)$, Bayes just* rescales *the numerators to sum to 1 — hence:*

$$\text{posterior} \;\propto\; \text{prior} \times \text{likelihood}$$

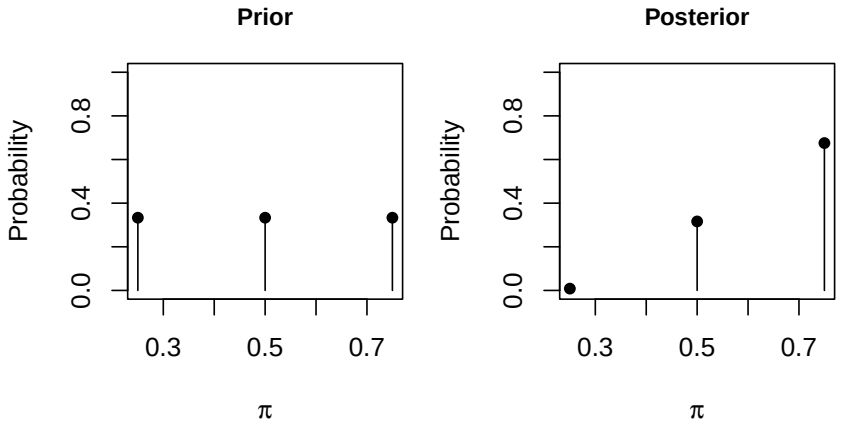
Coin Flips: Three Hypotheses

Flip a coin 10 times, get 7 heads. Frequentist: how likely is this data if the coin is fair? Bayesian: what is the probability of each type of coin, given this data?

Take three hypotheses — $\pi = .25, .50, .75$ — with equal priors. The likelihood is just the binomial.

```
pi <- c(.25, .50, .75)
prior <- rep(1/3, 3)
post <- dbinom(7, 10, pi) * prior     # likelihood x prior
# divide by the sum of numerators
post.norm <- post / sum(post)
round(post.norm, 4)
```

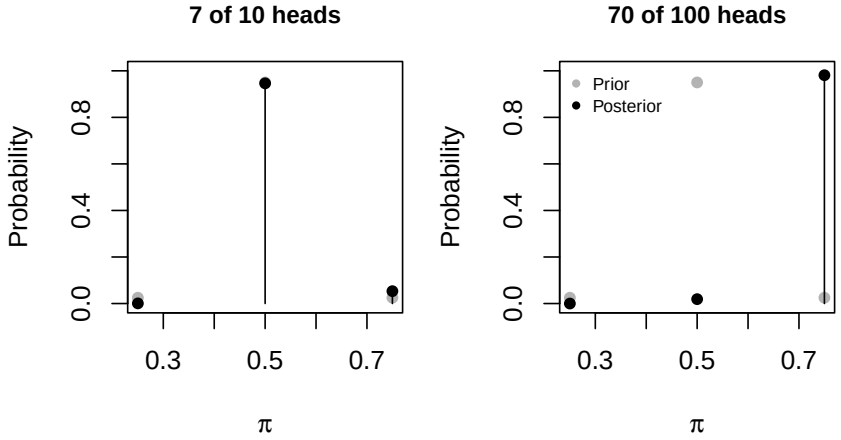
```
[1] 0.0083 0.3162 0.6754
```



Posterior: under 1% that this is a 25% coin, ~30% that it is fair, ~67% that it is a 75% coin.

Priors matter — and data can overwhelm them

Equal priors were silly: unfair coins are rare. Try a prior of 95% fair.



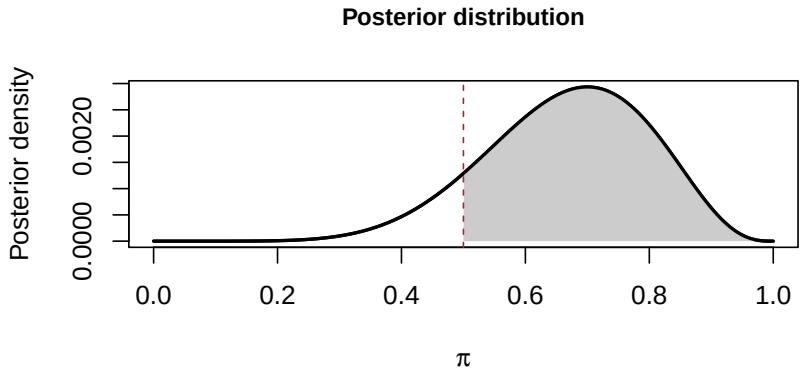
With 7/10 the strong prior barely moves. With 70/100 — the same proportion, *ten times the data — the evidence overwhelms the prior.*

Scaling Up to a Posterior Distribution

Nothing stops us using 1,001 hypotheses instead of 3.

```r
pi <- seq(0, 1, .001)
prior <- rep(1/1001, 1001)
post.norm <- dbinom(7, 10, pi) * prior
post.norm <- post.norm / sum(post.norm)
sum(post.norm[pi > .5])  # P(coin favors heads | data)
```

```
[1] 0.8860734
```



This is the payoff. *Frequentist statistics only lets us reject or retain a null. Here we have a full distribution over hypotheses, so we can make direct probability claims — an 88.6% chance this coin favors heads.*

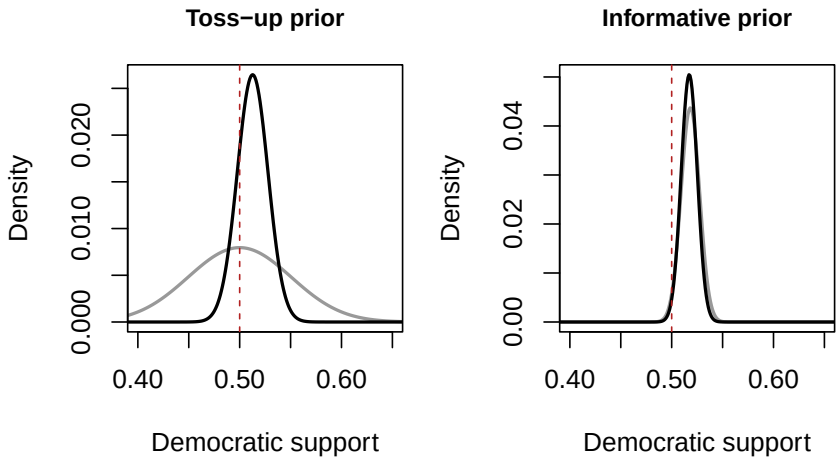
The Bayesian analogue of a confidence interval is a **credible interval**: *the range containing 95% of posterior probability. And unlike a confidence interval, you* may *say "there is a 95% probability the parameter is in here."*

Real Application: A Poll

A poll of 1,000 has 514 supporting the Democrat. A `t.test` *against* $\mu = .5$ *gives* $p \approx .38$ *— we fail to reject a tied race, and that is all we can say.*

Bayesian instead: put a prior on the race (a beta distribution assigns density between 0 and 1), then update.

```
Toss-up CI: [0.482, 0.542]
```




```
Informative CI: [0.501, 0.533]
```

With an informative prior built from three previous polls, the credible interval is narrower *from the same data — prior information is doing real work.*

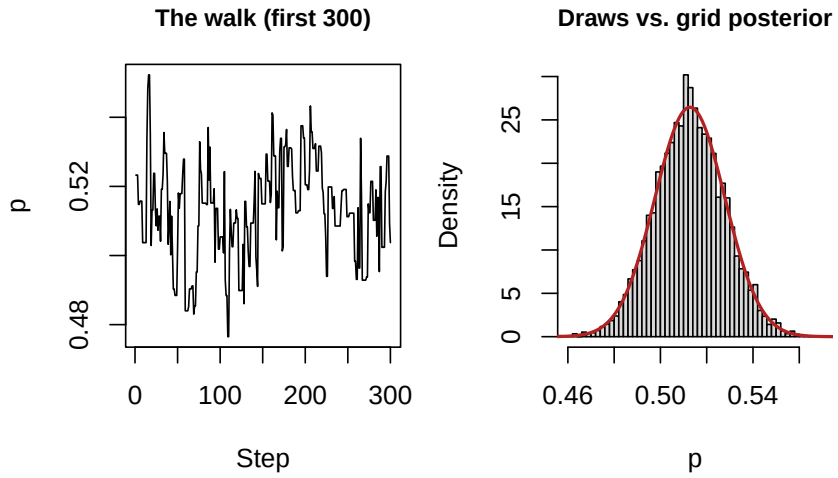
Getting the Posterior by Sampling

Enumerating every hypothesis worked fine for one parameter. But a regression needs a posterior for every combination *of α and β; with 7 parameters and 100 candidate values each, that is 100^7 points. Hopeless.*

Instead we **sample** *the posterior with a random walk — the* **Metropolis algorithm**:

1. *Start at some parameter value; compute the posterior height there.*
2. *Propose a nearby value (a small random step).*
3. *Compute the posterior there. If* higher, *move. If* lower, *move only sometimes — with probability equal to the ratio of heights.*
4. *Record where you stand. Repeat.*

The walk spends the most time where the posterior is highest.



The histogram of sampled values traces the exact curve the grid gave us — computed by wandering, not enumerating.

```
mean(p.draws)                    #posterior mean
```

```
[1] 0.5126977
```

```
quantile(p.draws, c(.025, .975))  #95% credible interval
```

```
     2.5%     97.5%
0.4836261 0.5422909
```

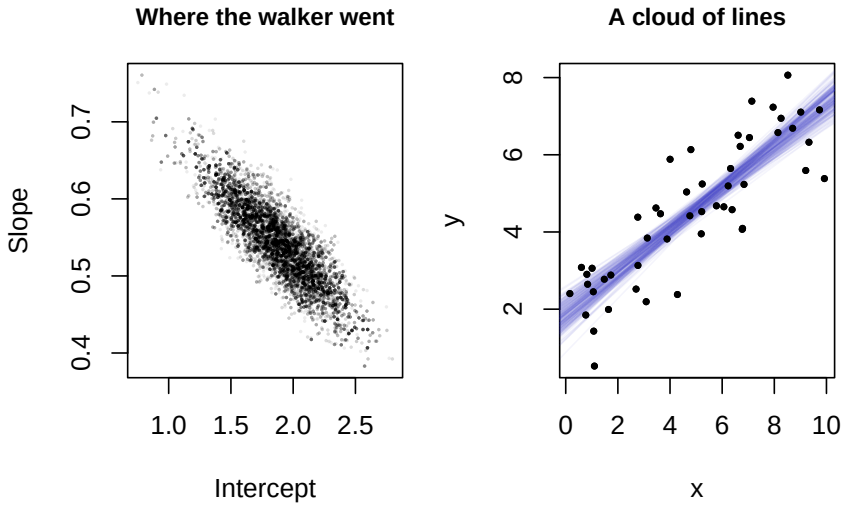
```
mean(p.draws > 0.5)              #P(Democrat ahead)
```

```
[1] 0.8002778
```

Summarize the draws exactly as you would bootstrap resamples.

Two Parameters: The Posterior Is a Cloud of Lines

Same algorithm — the walker now roams a plane *of* (α, β) *pairs instead of a line. Each draw is a pair, which is to say* each draw is a whole line.



The cloud leans on a diagonal: a steeper slope must pair with a lower intercept to keep the line through the data, so the two are not independent.

Where the lines bunch, the posterior is confident; where they fan, it is uncertain. That fan is the Bayesian answer to "what is the relationship between x and y" — not one best line, but the full set of plausible lines.

```
colMeans(draws)  #posterior means
```

```
    alpha      beta
1.8745046 0.5429588
```

```
#95% CI for the slope
quantile(draws[, "beta"], c(.025, .975))
```

```
     2.5%     97.5%
0.4437061 0.6437415
```

```
mean(draws[, "beta"] > 0)  #P(slope is positive)
```

```
[1] 1
```

```
confint(lm(y ~ x))  #least squares, for comparison
```

```
                2.5 %   97.5 %
(Intercept) 1.3682705 2.432993
x           0.4453066 0.634123
```

Essentially the same interval as OLS. Adding more parameters changes nothing conceptually — each step just proposes a move in more dimensions. This is exactly what Stan, brms, and PyMC do.

Why We Teach Frequentist Statistics

1. **The social sciences run on it.** *You cannot read the literature without p-values and confidence intervals.*
2. **Bayesian computation used to be intractable.** *The frequentist framework was built to be doable with pencil and paper.*
3. **The prior is genuinely controversial.** *Two researchers with different priors can reach different conclusions from identical data.*
4. **With enough data the two agree.** *At the sample sizes we have worked with all semester, the answers are nearly identical. The differences bite with small samples or strong prior information.*

Frequentism gives a rigorous, prior-free way to evaluate evidence. Bayes gives a principled way to fold in prior knowledge and make direct probability statements. Treat them as complements — but if you only know one, know the frequentist framework.

Multivariate Regression

PSCI1801 Chapter 12

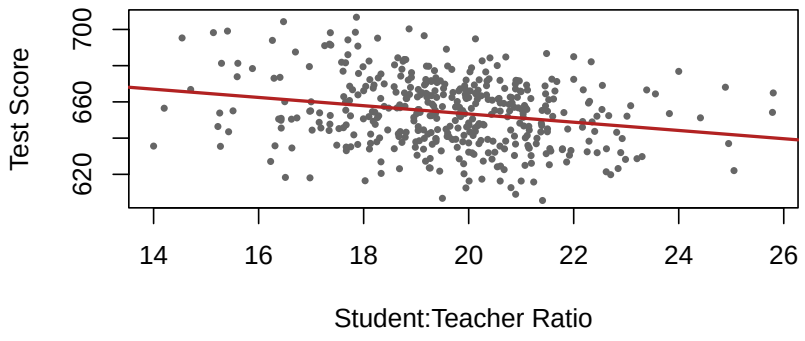
This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Regression with More Than One Independent Variable

The key feature of OLS: we can include multiple independent variables in one model.

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$$

Running example: California schools. Does the student/teacher ratio (STR) predict test scores?



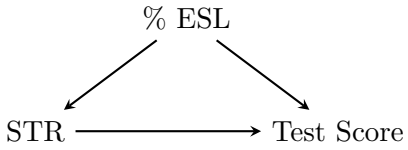
	Est	SE	t	p
(Intercept)	698.933	9.467	73.82	<0.001
STR	-2.280	0.480	-4.75	<0.001

Bivariate result: one additional student per teacher is associated with a **2.3 point drop** *in test scores. The intercept (698.9) is the average score when $STR = 0$ — mechanically necessary, substantively meaningless.*

Omitted Variable Bias

"Correlation does not equal causation" is true but lazy. The useful question: is some third variable causing both *x and y?*

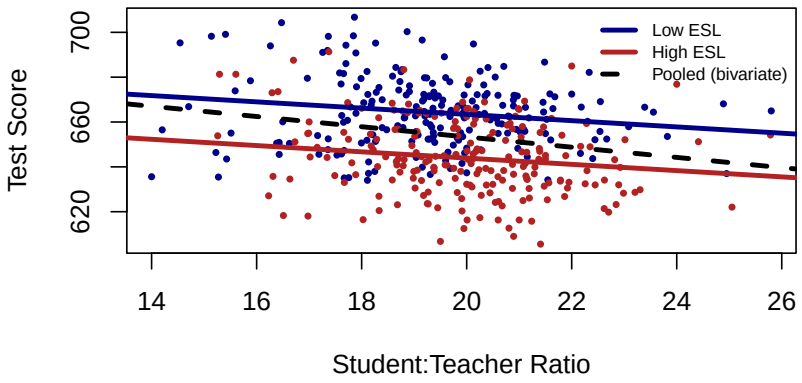
We are looking for a **common cause** *—* a confounder.



Schools with more ESL students may have both higher STR (poorer tax base → fewer teachers) and lower test scores (tests under-measure non-native speakers). The path STR ← ESL → Test is a **backdoor**; *controlling for ESL closes it.*

Classic absurd examples: firefighters at a fire and property damage (both caused by fire severity); ice cream sales and violent crime (both caused by temperature).

Seeing it: split the sample by ESL



Mechanically, regression now chooses **one slope and two intercepts** *such that the sum of squared residuals is minimized. The dashed pooled line is* steeper *than the within-group slope.*

	Est	SE	t	p
(Intercept)	691.327	8.131	85.02	<0.001
STR	-1.396	0.417	-3.35	<0.001
high.esl	-19.491	1.576	-12.37	<0.001

The STR coefficient shrinks from -2.3 to -1.4. Part of the original association was really ESL composition. The gap between the two lines is the `high.esl` *coefficient (≈ 19.5 points).*

Interpreting Multivariate Coefficients

The magic phrase: **"holding the other variables constant."**

$\hat{\beta}_j =$ change in y for a 1-unit change in x_j, holding all other xs constant

Same intercept rule, generalized:

$$\hat{\alpha} = \text{average } y \text{ when } \textbf{all} \text{ predictors} = 0$$

This never changes, no matter how many variables you add or how they are measured. Everything else is bookkeeping.

Why one slope and two intercepts, rather than letting the slope vary? Letting slopes differ is a separate (second-order) question requiring an **interaction**:

$$\text{score} = \alpha + \beta_1 STR + \beta_2 ESL + \beta_3 (STR \times ESL)$$

Covered in the Interaction and Prediction chapter.

Multiple Continuous Variables

Instead of a binary split, use `english` *in its original continuous form. With two continuous predictors the regression* line *becomes a* **plane**: *choose an intercept and two slopes minimizing vertical distance to every point.*

(See the interactive 3D version of this plot in the textbook — the points form a cloud in three dimensions and regression fits a flat plane through them.)

	Est	SE	t	p
(Intercept)	686.032	7.411	92.57	<0.001
STR	-1.101	0.380	-2.90	0.004
english	-0.650	0.039	-16.52	<0.001

- *Holding % ESL constant:* $+1$ *student per teacher* $\rightarrow -1.1$ *points.*
- *Holding STR constant:* $+1$ *percentage point ESL* $\rightarrow -0.65$ *points.*
- *Intercept 686: average score with no students and all English speakers. Still not helpful.*

Add a third predictor and visualization becomes impossible. From here on, reason directly from the coefficients.

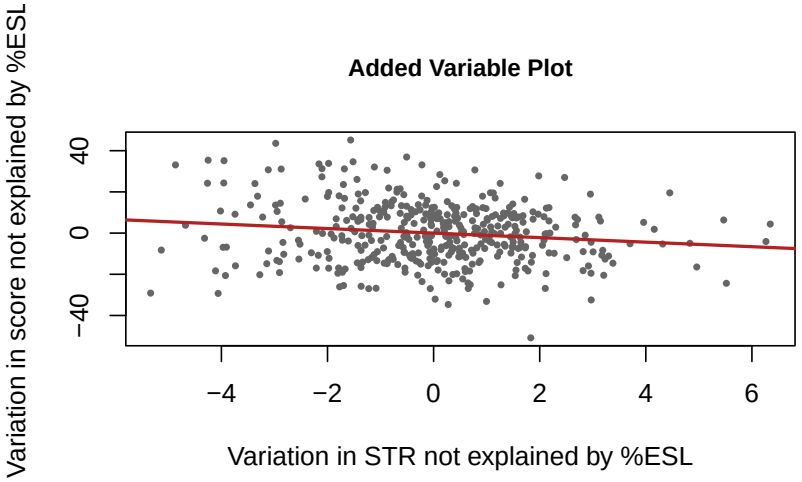
What "Holding Constant" Actually Does

We can rebuild a multivariate coefficient out of bivariate pieces. (Intuition only — you never need to do this to run a regression.)

Step 1. *Regress STR on %ESL, keep the residuals: the variation in STR* not *explained by %ESL.*

Step 2. *Regress score on %ESL, keep the residuals: the variation in score* not *explained by %ESL.*

Step 3. *Regress the leftovers on the leftovers.*



```
added-variable slope       full model STR
             -1.1013                -1.1013
```

The slopes are identical. **"Controlling for %ESL" means stripping the %ESL-related variation out of both variables and looking at the relationship in what is left.**

Many Variables at Once

Just keep adding to the right-hand side. Order does not matter — it is only addition.

```
                Est    SE      t      p
(Intercept) 675.608 5.309 127.26 <0.001
STR          -0.560 0.229  -2.45  0.015
english      -0.194 0.031  -6.19 <0.001
lunch        -0.396 0.027 -14.46 <0.001
income        0.675 0.083   8.10 <0.001
```

Every coefficient is still a partial effect. The intercept becomes increasingly absurd: average score for a school with no students, no ESL students, no subsidized lunches, and zero income $= 675$*. Ok!*

Statistical vs. substantive significance

All four variables are "statistically significant" — we are confident the effect is not zero. That says **nothing** *about importance.*

- *Do* **not** *use p-values to rank importance. Treat statistical significance as a yes/no and nothing more.*
- *Do* **not** *use raw coefficient size either — "one unit" differs across variables.*

Demonstration: `income` *is measured in thousands of dollars, with coefficient* 0.675*. Re-express it in dollars:*

```
income in $1000s     income in $1
         0.67498          0.00067
```

```
[1] TRUE
```

The coefficient becomes 0.00067 *and* **every other coefficient is unchanged**. *Income did not become 1000 times less important; "one unit" changed meaning.*

Standardizing for a like-to-like comparison

Put every variable on a common scale — how much does a one standard deviation *shift matter?*

$$z_i = \frac{x_i - \bar{x}}{sd(x)} \qquad \Rightarrow \qquad \text{mean} = 0, \;\; sd = 1$$

```
(Intercept)     english      income         STR
    654.157      -3.553       4.877      -1.060
      lunch
    -10.751
```

Now one unit = one standard deviation for every predictor, so coefficients are directly comparable. A one-SD shift in free lunch *moves scores by nearly* **11 points**; *a one-SD shift in* STR *by about* **1 point**. *Material conditions dominate here.*

Bonus: *standardizing makes every mean zero, so the intercept is now the average y when all predictors are* at their means *— an actually useful number (*654*), not a school that cannot exist.*

What Variables to Add?

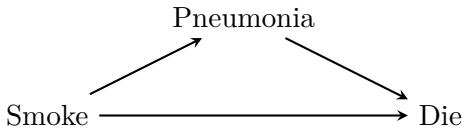
This is where you can get yourself in trouble. Specifying a regression is hard, theoretical work, and more variables is not better. *Three traps: two about bias, one about variance.*

Trap 1: Post-treatment bias (a mediator)

Simulated: smoking causes pneumonia, pneumonia causes death, smoking has no *direct effect.*

```
                         (Intercept)  smoke
die ~ smoke                   0.0744 0.1503     NA
die ~ smoke + pneumonia       0.0496 0.0000 0.4967
```

Smoking raises death probability by 15 points — until we "control for" pneumonia, at which point the effect goes to **zero**. *Smoking kills people* by giving them pneumonia. *Controlling for the mediator removes the very effect we want.*

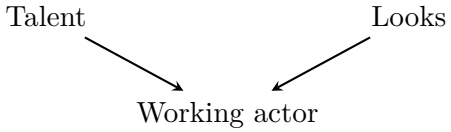


Note the arrow directions versus the confounder diagram — that is the whole difference. Another example: the effect of gender on wages "controlling for" length of parental leave controls away one of the reasons the gap exists.

Trap 2: Collider bias

A **collider** *is a common* effect *of two variables. Conditioning on one* creates *a relationship that isn't there.*

```
overall among actors
 -0.012       -0.656
```

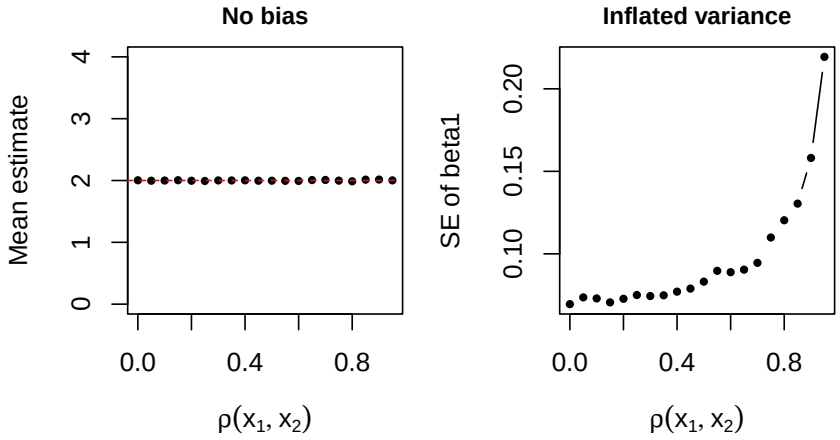


Talent and looks are unrelated in the population. But you need some combination *of the two to make it, so among working actors the ones who can't act got there on looks and vice versa — a spurious* negative *correlation.*

Same logic: GRE scores vs. research ability among PhD admits; height vs. scoring in the NBA. The general phenomenon is **Berkson's Paradox**.

Trap 3: Multicollinearity (a variance problem)

Adding a control highly correlated with your predictor of interest does not *bias anything — it inflates the standard error.*



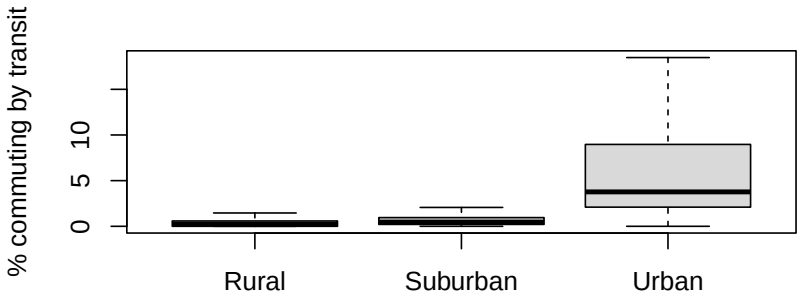
The estimate stays centered on the truth (2); the standard error roughly **triples** *(*$0.07 \to 0.22$*). When* x_1 *and* x_2 *move together the data cannot cleanly separate their effects.*

Summary of the three

- **Confounder** *(common cause of x and y):* **control for it.**
- **Mediator** *(on the path from x to y):* **don't** *— that's post-treatment bias.*
- **Collider** *(common effect of x and y):* **never** *— it invents a relationship.*

Regression with Categorical Variables

Suppose density is only recorded as rural (1) / suburban (2) / urban (3).



Treating it as continuous forces the two jumps (rural→suburban, suburban→urban) to be equal. *The boxplot says they clearly are not — urban is the outlier. And for an* un-ordered *variable (census region) the approach collapses entirely.*

Dummy variables

Create one indicator per category. They are mutually exclusive and exhaustive, so all three cannot be in the model: *if every indicator were 0 the intercept would be undefined. R drops one automatically.*

```
(Intercept)    ruralTRUE suburbanTRUE
      8.150       -7.666       -7.344
```

$$\widehat{\text{transit}} = 8.15 - 7.66 \cdot \text{Rural} - 7.34 \cdot \text{Suburban}$$

The omitted category is the **reference**. *Here rural = suburban = 0 means* urban, *so:*

- *Intercept = mean transit commuting in* urban *counties.*
- *Each coefficient = difference in means between that category and* urban.
- *Predicted rural value* $= 8.15 - 7.67 = 0.48$, *exactly the rural group mean.*

Change the reference category and the numbers change but the relationships are identical.

Factor variables

Rather than building indicators by hand, let R do it:

```
acs$census.region <- as.factor(acs$census.region)
acs$census.region <- relevel(acs$census.region,
                             ref = "northeast")
m7 <- lm(median.income ~ percent.transit.commute
         + census.region,
         data = acs)
```

	Est	SE	t	p
(Intercept)	57840.621	904.668	63.94	<0.001
percent.transit.commute	1049.999	74.528	14.09	<0.001
census.regionmidwest	-4995.192	971.664	-5.14	<0.001
census.regionsouth	-11203.467	950.651	-11.79	<0.001
census.regionwest	-3779.474	1058.925	-3.57	<0.001

`as.factor()` tells R to expand the variable into dummies automatically; `relevel()` chooses the reference category. Each region coefficient is now a difference from the northeast, holding transit commuting constant.

Tying It All Together

The interpretation never changes. *Continuous, binary, or a set of dummies — every coefficient is the change in y for a one-unit change in that variable, holding all others constant; the intercept is the average y when all predictors are zero. If you can say those two sentences and be specific about what "one unit" and "zero" mean, you can interpret any model.*

The hard part is specification. *R will happily run any regression you ask it to. Deciding what belongs on the right-hand side — confounder yes, mediator no, collider never, and watch the standard errors — is the judgment that separates a convincing analysis from a misleading one.*

Regression Functional Form

PSCI1801 Chapter 14

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Regression as a Data-Generating Process

You can run OLS on any numeric variables — but that does not mean it gives the right *answer. Start by rewriting the regression equation three ways:*

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_i \qquad \text{(the fitted line)}$$

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i \qquad \text{(actual data = line + residual)}$$

$$Y_i = \alpha + \beta X_i + \epsilon_i \qquad \text{(the population DGP)}$$

Drop the hats for the population and rename the residual $\hat{u}$ to the error ϵ. This third equation says: y's are produced by x's through a slope and intercept, plus irreducible error.

For OLS to recover the truth, five **Gauss-Markov assumptions** *must hold:*

1. *X has full rank.*
2. *The model is correctly specified: $Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$.*
3. *$E[\epsilon^2 \mid X] = \sigma^2$ (homoskedasticity).*
4. *$E[\epsilon_i \epsilon_j \mid X] = 0$ for $i \neq j$ (no autocorrelation).*
5. *$E[\epsilon_i \mid X] = 0$ (no confounding).*

GM1: Full Rank

No variable may be a linear combination of others. We already met this: a complete set of category dummies forces R to drop one. Same

problem if a variable has no variation *— a constant is a linear transform of the intercept.*

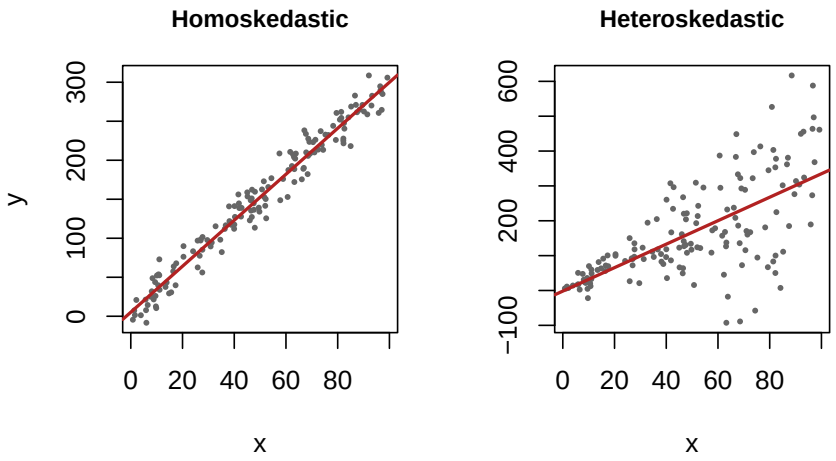
You never have to police this one. R simply returns `NA` *and moves on.*

GM5: No Confounding

This is the causality assumption — that $\hat{\beta}$ *is the* true effect *of* X *on* Y*. It holds when the errors are uncorrelated with anything else that matters, i.e.* there are no confounding variables. *This is the omitted-variable-bias material from the multivariate chapter.*

GM3: Homoskedasticity

The error variance must be constant across levels of X*. Compare:*



The right panel fans out. The consequence is **not** *bias in* $\hat{\beta}$ *— it is that the reported standard error is* wrong:

```
betas <- rep(NA,10000)
ses <- rep(NA,10000)

for(i in 1:10000){
x <- runif(100, min=0, max=100)
y <- 2 + 3*x + rnorm(100, mean=0, sd=x*2)
m <- lm(y ~ x)

betas[i] <- coef(m)["x"]
```

```
ses[i] <- sqrt(vcov(m)["x","x"])
}

#Empirical standard error
sd(betas)
```

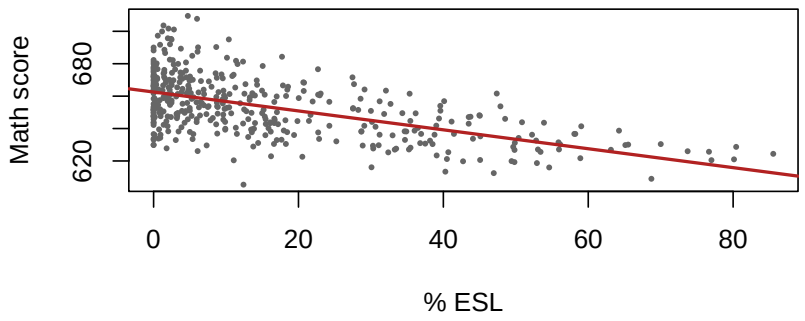
```
[1] 0.4454989
```

```
#Average calculated standard errors in samples
mean(ses)
```

```
[1] 0.4009148
```

Diagnosing and fixing it

Think theoretically first. Do we expect more variation in math scores at low or high ESL? Schools with few ESL students sit in all sorts of neighborhoods with all sorts of funding; high-ESL schools are more alike. So expect more spread at the low end.



The formal test is Breusch-Pagan. A significant result rejects the null of homoskedasticity:

```
m <- lm(math ~ esl, data = dat)
bptest(m)
```

```
    studentized Breusch-Pagan test
```

```
data:  m
BP = 12.084, df = 1, p-value = 0.0005085
```

Fix 1 — model it better. *The extra variance is unmodeled heterogeneity. Add variables that explain those schools:*

```
m.full <- lm(math ~ esl + lunch + calworks + income,
             data = dat)
bptest(m.full)
```

```
    studentized Breusch-Pagan test

data:  m.full
BP = 7.4826, df = 4, p-value = 0.1125
```

No longer heteroskedastic.

Fix 2 — robust standard errors. *Huber-White ("robust") SEs from the* `sandwich` *package:*

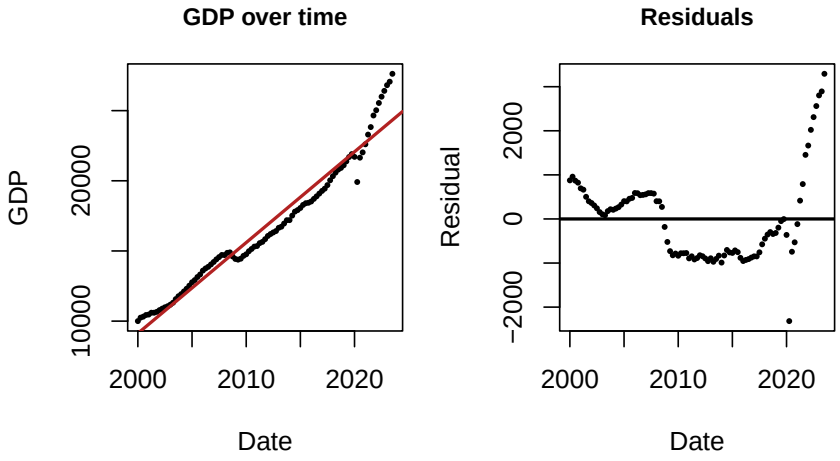
```
coeftest(m, vcov = vcovHC(m, type = "HC1"))
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 662.539315   1.046782 632.929 < 2.2e-16 ***
esl          -0.583245   0.033671 -17.322 < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GM4: No Autocorrelation

Errors must be independent across observations. Two things usually break this: **time** *and* **clusters**.

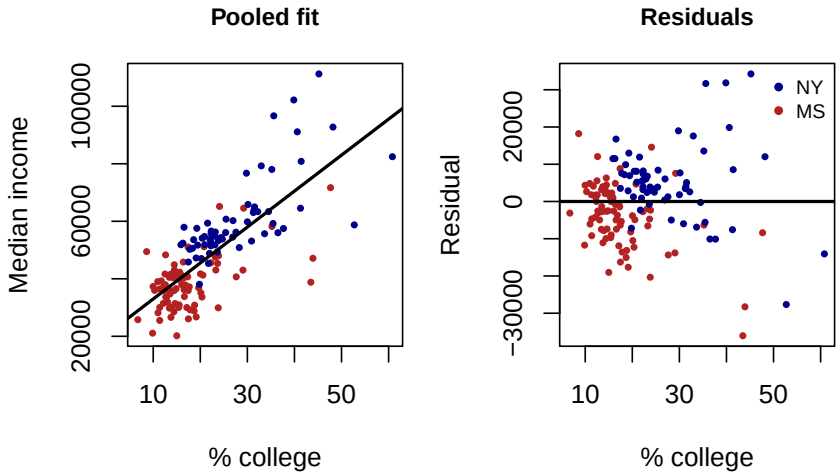
Time



Tell me the residual at one quarter and I can guess its neighbour's. That is exactly what independence forbids. (Time-series methods are their own course — for now, just recognize the pattern.)

Clusters

County income vs. college share, in NY and MS only:



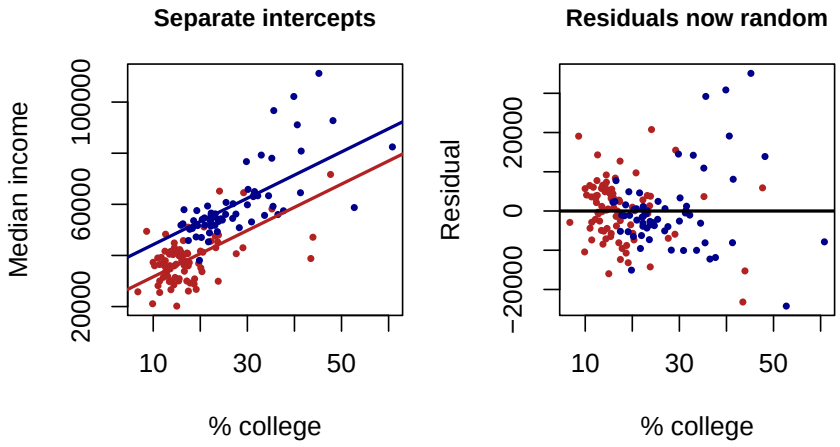
Mississippi counties sit below *the line; New York counties sit* above *it. Within clusters, errors are correlated.*

Fix 1 — clustered standard errors *(also heteroskedasticity-robust, a bonus):*

```r
coeftest(m1, vcov = vcovCL(m1, cluster = ~ state.abbr,
                           type = "HC1"))
```

```
t test of coefficients:

                Estimate Std. Error t value  Pr(>|t|)
(Intercept)     20468.08    1048.71  19.517 < 2.2e-
16 ***
percent.college  1250.74     154.34   8.104 2.223e-
13 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

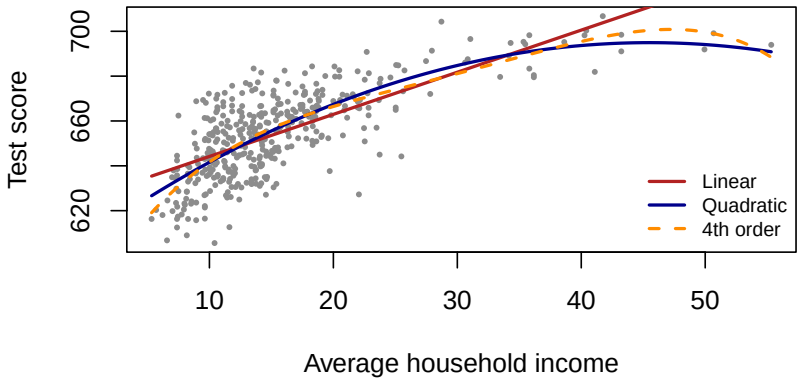
Fix 2 — model the clusters. *Add an indicator, which we already know how to read: two intercepts, one slope.*



Scaling this to all 50 states means one indicator per state — a **fixed effects** *regression. Include* `factor(state.abbr)`*, or use* `felm()` *from the* `lfe` *package to suppress the unreadable dummy output. Best practice is to use fixed effects* and *cluster the standard errors.*

GM2: Correct Functional Form

The model you specified must be the model that generated the data. The usual violations: unmodeled non-linearity *and unmodeled* interactions.



The straight line clearly misfits: scores rise steeply from income 5 to 25, then flatten. The fix is a **polynomial** *— which simply lets the line curve:*

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon$$

```
dat$income2 <- dat$income^2
round(coef(lm(score ~ income + income2, data = dat)), 4)
```

```
(Intercept)      income     income2
   607.3017      3.8510     -0.0423
```

Interpreting a polynomial: the effect now depends on x

With a curve there is no single "effect of income." Use the power rule *to get the slope at any point:*

$$\frac{\partial y}{\partial x}(aX^k) = k\,aX^{\,k-1} \qquad \frac{\partial y}{\partial x}(a) = 0$$

For an ordinary regression this recovers exactly what we already knew:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \quad \Rightarrow \quad \frac{\partial y}{\partial x} = \hat{\beta}$$

With a squared term:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \quad \Rightarrow \quad \frac{\partial y}{\partial x} = \hat{\beta}_1 + 2\hat{\beta}_2 x$$

```
m <- lm(score ~ income + income2, data=dat)
coef(m)[2] + 2*coef(m)[3]*20
```

```
  income
2.158656
```

```
coef(m)[2] + 2*coef(m)[3]*30
```

```
  income
1.312487
```

```
coef(m)[2] + 2*coef(m)[3]*50
```

```
    income
-0.3798508
```

The effect of income shrinks as income rises — exactly what the picture showed.

Don't over-fit

A 4th-order polynomial fits even better, and R^2 climbs. Why not go further?

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + 4\beta_4 x^3$$

Because the goal is to describe the population *DGP, not to trace every wiggle of* this sample. *Taken to the extreme you would fit the noise perfectly and generalize to nothing. The 4th-order fit above is over-fitting. Balance a plausible functional form against over-fitting — the same tension we saw when we declined to connect the group means back in the regression chapter.*

This chapter was taught in previous sessions of the course. In F2026 we will likely not have time to cover it. If we end up covering this material it will become testable. Until then, consider this extra reference material.

Regression Interaction and Prediction

PSCI1801 Chapter 15 (supplementary)

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

Interaction

A polynomial let the effect of a variable change over its own level. *An* **interaction** *lets the effect of a variable change* across the levels of a second variable.

Question: does the effect of household income on test scores differ between K-6 and K-8 schools?

Adding an indicator gives two intercepts

This answers "do K-8 schools score differently?" — not "does income matter more *in one type?" For that we must* multiply *the variables.*

Multiplying gives two slopes

$$\widehat{score} = \alpha + \beta_1 Income + \beta_2 K8 + \beta_3 (Income \times K8)$$

```
m <- lm(score ~ income * k8, data = dat)
round(coef(m), 4)
```

```
(Intercept)      income          k8   income:k8
   625.0716      2.0132      0.6893     -0.1845
```

R creates all three terms automatically. Read them with the power rule — the derivative with respect to a variable is the effect of that variable:

$$\frac{\partial Score}{\partial Income} = \beta_1 + \beta_3 K8 \qquad \frac{\partial Score}{\partial K8} = \beta_2 + \beta_3 Income$$

The effect of each variable now depends on the level of the other. *(A polynomial is just a variable interacted with itself.)*

```
coef(m)["income"] + coef(m)["income:k8"]*0
```

```
  income
2.013211
```

```
coef(m)["income"] + coef(m)["income:k8"]*1
```

```
  income
1.828728
```

Two rules to memorize:

1. *The* **interaction coefficient is the difference between the two slopes**. *So its hypothesis test asks whether the effect* changes *across levels of the other variable.*
2. *The coefficient on a* constituent term *(`income`) is the effect of that variable* when the other equals zero. *ALWAYS. This often makes it a nonsense number.*

Continuous × Continuous Interaction

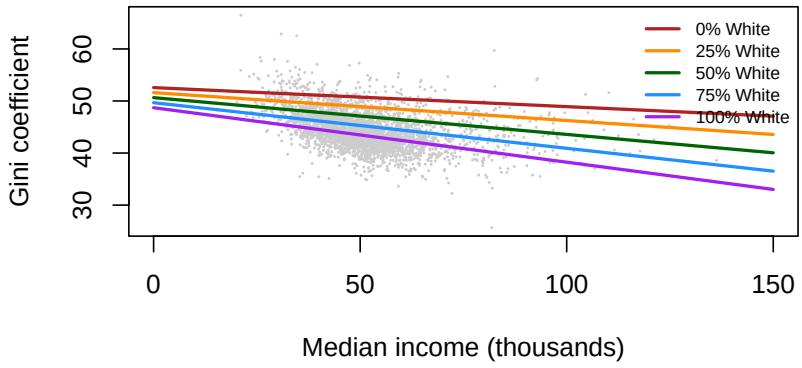
County median income vs. income inequality (Gini), interacted with percent white.

```
m4 <- lm(gini ~ median.income * percent.white, data = acs)
round(coef(m4), 5)
```

```
                (Intercept)               median.income
                   52.57554                    -0.03660
             percent.white median.income:percent.white
                   -0.03866                    -0.00068
```

$$\frac{\partial Gini}{\partial Income} = \beta_1 + \beta_3 \cdot PercWhite$$

Now the "one unit change in income" differs at every *level of percent white — each additional point of percent white changes the* effect *of income by* β_3.



The marginal effect plot

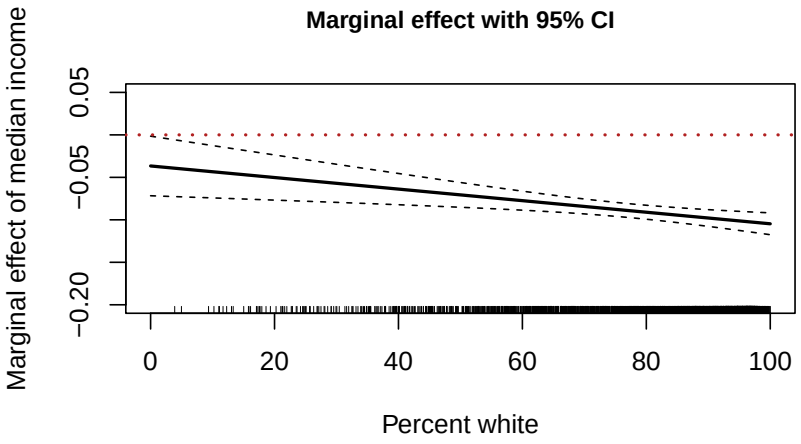
Usually more useful than a family of lines: plot the effect itself *across levels of the moderator.*

To add a confidence band we need the variance of a sum, *which is not just the sum of variances:*

$$V(X+Y) = V(X) + V(Y) + 2\,\mathrm{Cov}(X,Y)$$

$$V(\beta_1 + \beta_3 W) = V(\beta_1) + W^2 V(\beta_3) + 2W\,\mathrm{Cov}(\beta_1, \beta_3)$$

The coefficients do *covary, so that last term matters. Get all the pieces from the variance-covariance matrix,* `vcov(m)` *— variances on the diagonal, covariances off it. Note W appears in the formula, so* the standard error changes across levels of the moderator.



The **rug** *at the bottom shows where counties actually are — always add it. A marginal effect estimated in a region with almost no data should not persuade you.*

Prediction

So far regression has been a tool of **explanation** *("how does x affect y?"). The other use is* **prediction** *("given this relationship, what do we expect from* new *data?"). None of the math changes.*

Since $\hat{y} = \hat{\alpha} + \hat{\beta}x$, we can plug in any x we like — including values we never observed.

```
round(coef(lm(seconds ~ year, data = mwr)), 3)
```

```
(Intercept)        year
  54898.384     -23.755
```

Each year the marathon world record falls by about 24 seconds. But a straight line fits these data badly — a 3rd-order polynomial is better.

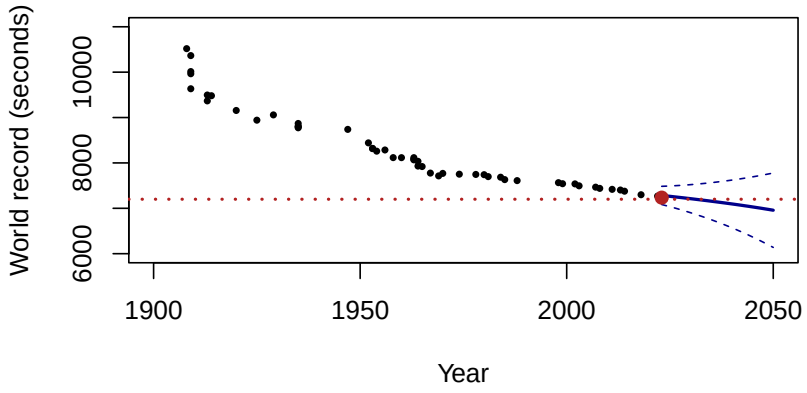
Doing the confidence band by hand uses the same variance-of-a-sum formula:

$$V(\alpha + \beta \cdot year) = V(\alpha) + year^2 V(\beta) + 2\, year\, \mathrm{Cov}(\alpha, \beta)$$

Mercifully, `predict()` does all of this for us — and scales to models where the hand calculation would be brutal:

```
m2 <- lm(seconds ~ poly(year, 3), data = mwr)
fut <- seq(2023, 2050, 1)
head(predict(m2, newdata = data.frame(year = fut),
             interval = "confidence"), 3)
```

```
       fit      lwr      upr
1 7283.737 7083.140 7484.333
2 7273.785 7059.807 7487.764
3 7263.741 7035.667 7491.815
```



Prediction: the two-hour barrier (dotted line) falls around 2030, with wide uncertainty. The red point is Kelvin Kiptum's 2:00:35 in 2023 — well off the line.

The limit of prediction: *we can only use what we have, and must assume current trends continue. We cannot model in future intercept shifts — new shoe technology, training methods, nutrition (or doping).*

Out-of-Sample Prediction and Over-Fitting

Build a **likely voter model**: *predict who actually turns out. Train on 2016 and 2018 CCES data, test on 2020.*

```
cces$vote.intent <- relevel(factor(cces$vote.intent),
                            ref = "No")
train <- cces[cces$year %in% c(2016, 2018), ]
test  <- cces[cces$year == 2020, ]

m1 <- lm(vv.turnout ~ vote.intent, data = train)
m2 <- lm(vv.turnout ~ vote.intent + prim.turnout +
           educ:white, data = train)
m3 <- lm(vv.turnout ~ female + vote.intent +
           prim.turnout + educ:white:agegrp + pid3,
         data = train)
m4 <- lm(vv.turnout ~ pid3:vote.intent:prim.turnout +
           female:educ:white:agegrp, data = train)

err <- function(mod) {
  p <- predict(mod, newdata = test)
  mean((p > .5 & test$vv.turnout == 0) |
       (p < .5 & test$vv.turnout == 1),
       na.rm = TRUE)
}
round(c(simple = err(m1), medium = err(m2),
        complex = err(m3), kitchen.sink = err(m4)), 4)
```

```
      simple       medium      complex kitchen.sink
      0.2363       0.2115       0.1948       0.2333
```

Models 1 to 3 improve: more information about why *people vote gives better predictions. Then model 4 — interacting everything with everything — makes the classification error go* **up**.

This is over-fitting, and it is the same lesson as the polynomial. *The training data are a* sample, *carrying random noise. Estimating an interaction for every sub-category fits* that noise. *Past some point, a better fit in-sample means a* worse *fit out-of-sample.*

Optimizing a model for out-of-sample predictive accuracy is, essentially, **machine learning**. *There are fancier models and more details, but it is not a big leap from here.*

Where To Go From Here

Three directions out of this course:

1. **Pure statistics** — *more probability, more distributions, getting standard errors right. We scratched the surface.*
2. **Causal inference** — *an exciting, fast-moving field about when correlation* does *license causal claims.*
3. **Prediction** — *machine learning, baby.*