

# Descriptive Statistics

Gov 51: Data Analysis and Politics



Scott Cunningham

Harvard University

Week 2, Thursday

February 6, 2026

# A Simple Question

What does the typical American think about the president?

How do we answer this with data?

# Today's Roadmap

## 1. How We Summarize Data

- ▷ Measures of center (mean, median)
- ▷ Measures of spread (variance, standard deviation)

## 2. When Observations Count Differently

- ▷ Weighted statistics and why they matter

## 3. Why Pictures Matter

- ▷ Histograms, distributions, and telling stories with data



The Question: What Does “Typical” Mean?

# Setting Up the Problem

What does the typical American think about the president?

Gallup and other pollsters survey Americans constantly. But:

- ▷ Different states have different opinions
- ▷ Different states have different populations
- ▷ How do we summarize all this into one number?

Today we'll use state-level presidential approval data to learn how to summarize distributions.

# Loading Our Data

```
approval <- read.csv("state_approval.csv")
dim(approval)
## [1] 50 5

head(approval, 4)
##      state abbrev approval population  region
## 1 Alabama      AL       38    5024279  South
## 2 Alaska       AK       41     733391   West
## 3 Arizona      AZ       44    7151502   West
## 4 Arkansas     AR       36    3011524   South
```

50 states, with approval rating (%) and population.

Note: The R script on the course website has the full URL to load this data.

## First Look: The Raw Numbers

Here are approval ratings for a few states:

State	Approval (%)	Population (millions)
California	52	39.5
Texas	41	29.1
Wyoming	32	0.6
Vermont	57	0.6
Ohio	42	11.8

**What's the “typical” approval rating?**

Should Wyoming (0.6 million people) count the same as California (39.5 million)?



# Measures of Center



## Let's Start Simple

Before we use R, let's calculate by hand.

Here are approval ratings for 5 states:

**38, 41, 44, 36, 52**

**Calculate the mean:**

- 1.** Add them up:  $38 + 41 + 44 + 36 + 52 = 211$
- 2.** Divide by the count:  $211 \div 5 = 42.2$

The mean approval rating is 42.2%.

That's all the mean is: sum divided by count.

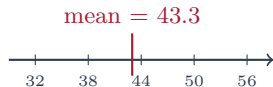
# The Mean: What R Does

The `mean()` function does exactly what we just did:

```
mean(approval$approval)
## [1] 43.26
```

For all 50 states, the mean approval is **43.3%**.

Simple, right? But there's a catch...



# The Median: The Middle Value

The **median** is the middle value when you sort the data.

```
# Sort the values
sort(approval$approval)
## [1] 32 34 34 35 35 36 36 37 37 38 38 38 39 39 39 39 40
## [18] 40 41 41 42 42 43 43 44 44 45 45 45 46 46 47 47 47
## [35] 48 48 49 49 50 50 50 51 52 52 53 56 57 58

# The median (middle value)
median(approval$approval)
## [1] 43.5
```

With 50 values, the median is the average of the 25th and 26th values.

Median = 43.5%

# Mean vs. Median: Why Both?

In our data: Mean = 43.3%, Median = 43.5% — almost the same! But that's not always true.

## The Mean

- ▷ Uses every value
- ▷ Sensitive to outliers
- ▷ Gets “pulled” by extreme values

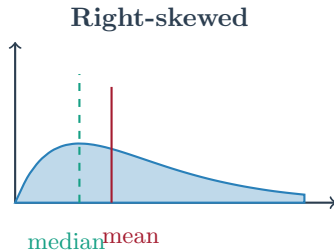
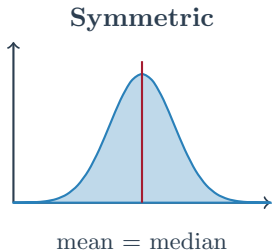
## The Median

- ▷ Only uses the middle
- ▷ Robust to outliers
- ▷ Ignores extreme values

Mean > median  $\Rightarrow$  right-skewed

| Mean < median  $\Rightarrow$  left-skewed

# Visualizing the Difference



Income is right-skewed: a few billionaires pull the mean above the median.

# Connection to Problem Set 1

In PS1, you'll calculate mean and median commute times.

**The question we'll ask:**

- ▷ Is the mean larger or smaller than the median?
- ▷ What does that tell you about the shape of the distribution?
- ▷ Does it match what you see in the histogram?

This is how you interpret data, not just calculate it.

## Now the Math

We've seen what the mean does. Here's the formula:

$$\text{Sample Mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- ▷  $x_i$  = each individual value
- ▷  $n$  = number of values
- ▷  $\sum$  = “add them all up”
- ▷  $\bar{x}$  = the mean (pronounced “x-bar”)

The formula just says: add up all the values, divide by how many there are.

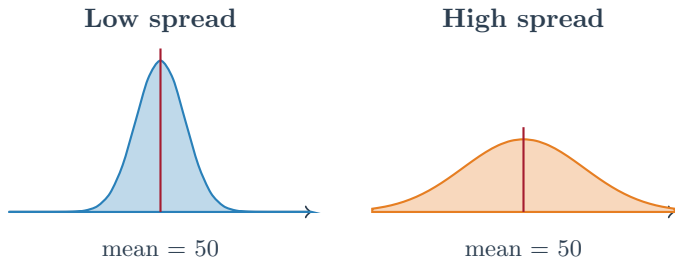


# Measures of Spread



# Center Isn't Everything

Two datasets can have the same mean but look completely different:



**We need measures of spread:** How much do the values vary around the center?

# The Range: Simplest Measure

The **range** is just max minus min:

```
min(approval$approval)
## [1] 32

max(approval$approval)
## [1] 58

range(approval$approval)
## [1] 32 58
```

Range =  $58 - 32 = 26$  percentage points

**Problem:** The range only uses two values. One outlier can make it huge.

# Percentiles: More Robust

Percentiles tell you where values fall in the distribution:

```
quantile(approval$approval)
##      0%    25%    50%    75%   100%
##      32     39     43     49     58
```

- ▷ 0th percentile (min): 32%
- ▷ 25th percentile (Q1): 39%
- ▷ 50th percentile (median): 43.5%
- ▷ 75th percentile (Q3): 49%
- ▷ 100th percentile (max): 58%

The middle 50% of states have approval between 39% and 49%.

## Getting Specific Percentiles

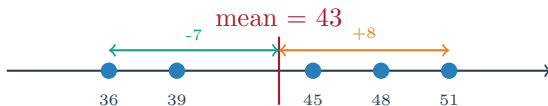
You can ask for any percentile:

```
# The 90th percentile  
quantile(approval$approval, 0.90)  
## 90%  
## 52  
  
# Multiple percentiles at once  
quantile(approval$approval, c(0.10, 0.50, 0.90))  
## 10% 50% 90%  
## 36 43 52
```

In PS1, you'll calculate the 90th percentile of commute times and write:  
"90% of commuters travel \_\_\_ minutes or less."

# Variance: Average Squared Distance from Mean

The **variance** measures how far values typically are from the mean.



## Steps:

1. Calculate each deviation:  $(x_i - \bar{x})$
2. Square them:  $(x_i - \bar{x})^2$
3. Average the squared deviations

# The Problem: Deviations Cancel Out

If we just add up the deviations from the mean:

- ▷ Some values are above the mean (positive deviation)
- ▷ Some values are below the mean (negative deviation)
- ▷ They cancel out:  $\sum(x_i - \bar{x}) = 0$

The sum of deviations from the mean is **always zero**.

So we can't just “average the deviations” to measure spread.

# The Solution: Square the Deviations

Square each deviation before summing:

- ▷ Squaring makes everything positive
- ▷ Bigger deviations get emphasized (squared distance)
- ▷ Now we can take a meaningful average

This gives us the **variance**: the average squared deviation from the mean.

There are other solutions (like absolute value), but squaring has nice mathematical properties.

## Variance in R

```
var(approval$approval)
## [1] 42.28
```

The variance is 42.28... but 42.28 *what?*

Units are “percent squared”—not very interpretable!

**Solution:** Take the square root to get back to original units.



# Variance Is Always Non-Negative

Look at the formula:

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

Each term  $(x_i - \bar{x})^2$  is a **squared number**.

- ▷ Squaring any real number gives something  $\geq 0$
- ▷ Sum of non-negative terms is non-negative
- ▷ Dividing by  $(n-1) > 0$  keeps it non-negative

**Result:** Variance  $\geq 0$ , always.

Variance = 0 only when every value equals the mean (no spread at all).

## Standard Deviation: Variance in Original Units

The **standard deviation** is the square root of variance:

```
sd(approval$approval)
## [1] 6.50
```

The standard deviation is 6.5 percentage points.

**Interpretation:** On average, states are about 6.5 percentage points away from the mean approval rating.

This is the measure of spread you'll report in your summary statistics tables.

# The Formulas

**Sample Variance:**  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

**Sample Standard Deviation:**  $s = \sqrt{s^2}$

Wait... why  $n - 1$  instead of  $n$ ?

# The $n - 1$ Question

**Quick intuition** (don't worry about the proof):

When we calculate variance, we first calculated the mean. That “used up” one piece of information.

- ▷ We have  $n$  data points
- ▷ But only  $n - 1$  independent pieces of information left
- ▷ This is called **degrees of freedom**

Dividing by  $n - 1$  instead of  $n$  corrects for this, giving us an unbiased estimate of the population variance.

R's `var()` and `sd()` functions use  $n - 1$  by default. That's what you want.

# Why Does Estimation “Use Up” Information?

**The problem:** We want to measure spread around the *true* mean  $\mu$ . But we don't know  $\mu$ —we estimated it with  $\bar{x}$ .

**The catch:**  $\bar{x}$  is the point that *minimizes* squared deviations.

- ▷ Deviations from  $\bar{x}$  are artificially small
- ▷ Dividing by  $n$  would systematically underestimate variance
- ▷ This is called **bias**

**The constraint:** Once you know  $\bar{x}$ , the deviations must sum to zero.

- ▷ If you know  $n - 1$  deviations, you can calculate the last one
- ▷ Only  $n - 1$  deviations are “free to vary”

Dividing by  $n - 1$  corrects for the bias. The estimate becomes **unbiased**.

## The summary() Shortcut

R's `summary()` function gives you many statistics at once:

```
summary(approval$approval)
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    32.00   39.00   43.50   43.26   49.00   58.00
```

This gives you:

- ▷ Min and Max (range)
- ▷ 1st and 3rd Quartiles (25th and 75th percentiles)
- ▷ Median (50th percentile)
- ▷ Mean

Note: `summary()` doesn't give you standard deviation—use `sd()` for that.

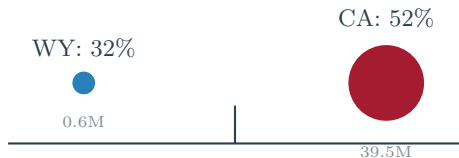


# Weighted Statistics

## Not All States Are Equal

Our mean approval (43.3%) treated every state equally.

But should Wyoming (576,000 people) count the same as California (39.5 million)?



If we want to know what the typical *American* thinks (not the typical *state*), California should count more.



# Weighted Mean

The **weighted mean** lets each observation count according to its weight:

```
# Unweighted mean (each state counts equally)
mean(approval$approval)
## [1] 43.26

# Weighted mean (weight by population)
weighted.mean(approval$approval, approval$population)
## [1] 44.52
```

- ▷ Unweighted: 43.3% (average across states)
- ▷ Weighted: 44.5% (average across people)

The weighted mean is higher because large, high-approval states (CA, NY) pull it up.

# The Weighted Mean Formula

$$\text{Weighted Mean: } \bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Where  $w_i$  is the weight for observation  $i$ .

**Intuition:** Instead of each value counting once, it counts  $w_i$  times.

When all weights are equal, this reduces to the regular mean.

# Connection to Problem Set 1

In PS1, you'll use American Community Survey (ACS) data.

The ACS has a variable called PERWT (person weight):

- ▷ Each person in the survey represents many people in the US
- ▷ PERWT tells you how many

**Two calculations:**

```
# Unweighted (average in the sample)
mean(commuters$TRANTIME)

# Weighted (average in the US population)
weighted.mean(commuters$TRANTIME, commuters$PERWT)
```

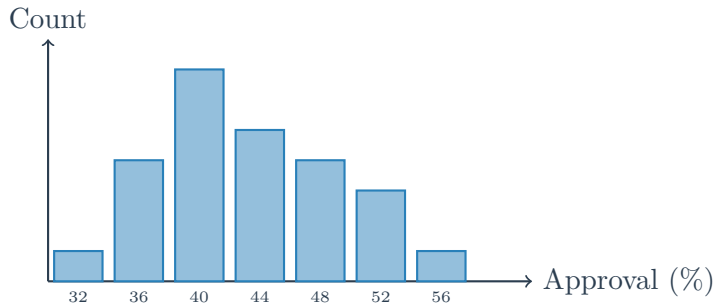
You'll compare these and explain what each number represents.



# Visualizing Distributions

# The Histogram

A **histogram** shows how values are distributed:



- ▷ X-axis: the variable (approval rating)
- ▷ Y-axis: how many observations fall in each bin
- ▷ **Shape** tells us about the distribution

# Creating a Histogram in R

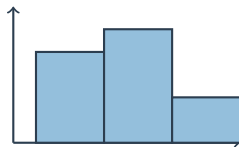
```
library(ggplot2)

ggplot(approval, aes(x = approval)) +
  geom_histogram(binwidth = 4,
                 fill = "steelblue",
                 color = "white") +
  labs(x = "Approval Rating (%)",
       y = "Number of States",
       title = "Distribution of Presidential Approval")
```

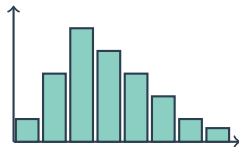
Key choices:

- ▷ **binwidth**: How wide is each bar? (Experiment!)
- ▷ **fill**: Color of the bars
- ▷ **color**: Color of the bar borders

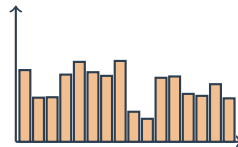
# Bin Width Matters



**Too few bins**  
Hides patterns



**Good**  
Shows shape



**Too many bins**  
Too noisy

There's no perfect answer—try different values and see what tells the clearest story.

# Describing Shape

When you look at a histogram, describe:

## Shape

- ▷ Symmetric?
- ▷ Right-skewed?
- ▷ Left-skewed?
- ▷ Bimodal?

## Center

- ▷ Where's the “middle”?
- ▷ Mean and median

## Spread

- ▷ How wide?
- ▷ Are values clustered or dispersed?

## Also note:

- ▷ Outliers (unusual values)
- ▷ Notable features (gaps, spikes)



## Example Description

Looking at our approval rating histogram:

*“The distribution of presidential approval across states is roughly symmetric, centered around 43%. Most states fall between 36% and 52% approval. There are no obvious outliers, though Vermont (57%) and Hawaii (58%) are notably high, while Wyoming (32%) and West Virginia (34%) are notably low.”*

**In PS1**, you'll write a similar description for commute times.



Putting It Together

# The Summary Statistics Table

In PS1, you'll create a table like this:

Variable	N	Mean	Std. Dev.	Min	Max
Approval (%)	50	43.3	6.5	32	58
Population (millions)	50	6.6	7.2	0.6	39.5

This table should be:

- ▷ Generated by code (not typed manually)
- ▷ Readable on its own (clear variable names, units)
- ▷ Rendered cleanly in your PDF

# Building the Table in R

```
stats <- data.frame(  
  Variable = c("Approval (%)", "Population (mil)"),  
  N        = c(length(approval$approval), length(approval$  
    population)),  
  Mean     = c(mean(approval$approval), mean(approval$population  
    )/1e6),  
  SD       = c(sd(approval$approval), sd(approval$population)/1  
    e6),  
  Min      = c(min(approval$approval), min(approval$population)/  
    1e6),  
  Max      = c(max(approval$approval), max(approval$population)/  
    1e6))  
knitr::kable(stats, digits = 1)
```

Full code in the R script on the course website.

# What You've Learned

## Concepts:

- ▷ Mean vs. median
- ▷ Skewness and shape
- ▷ Variance and standard deviation
- ▷ Degrees of freedom ( $n - 1$ )
- ▷ Weighted statistics

## R Functions:

- ▷ `mean()`, `median()`
- ▷ `min()`, `max()`, `range()`
- ▷ `var()`, `sd()`
- ▷ `quantile()`
- ▷ `weighted.mean()`
- ▷ `summary()`
- ▷ `ggplot() + geom_histogram()`



**Pictures Tell Stories**

# The Rhetoric of Quantitative Work

1. Beautiful pictures — patterns the eye can see
2. Beautiful tables — precise numbers for reference
3. Beautiful words — clear explanation, no confusion
4. Telling the truth well — integrity and craft

Our work is:

- ▷ Accurate and replicable (one button runs it all)
- ▷ Automated (figures and tables generated by code)
- ▷ Respectful of the reader's time

Pictures reveal patterns that tables hide.

# Visualization Principles: White Space

**White space** is not wasted space—it helps the reader.

## **Cluttered:**

- ▷ Every pixel filled
- ▷ Gridlines everywhere
- ▷ Labels on every point
- ▷ Reader overwhelmed

## **Clean:**

- ▷ Data stands out
- ▷ Minimal gridlines
- ▷ Labels where needed
- ▷ Pattern is clear

Edward Tufte: Maximize the “data-ink ratio”—every drop of ink should show data.



# Visualization Principles: Dual Y-Axes

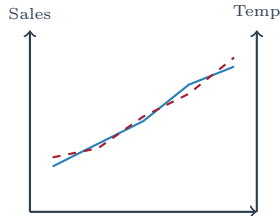
Dual y-axes are dangerous.

The problem:

- ▷ You control both scales
- ▷ You can make *any* two series look correlated
- ▷ Or make them look divergent
- ▷ Reader can't compare magnitudes

**Better alternatives:**

- ▷ Index to first period (= 100)
- ▷ Two separate panels (facets)



“Perfect correlation!”  
(or is it?)

# Case Study: Online Dating and the American Family

I'll show you pictures from my own research to illustrate these ideas.

## **The project:**

- ▷ Study of online dating's effect on marriage and fertility
- ▷ Data source: Craigslist Personals (now defunct)
- ▷ 166,000 personal ads recovered from Internet Archive
- ▷ Posts classified using GPT-4o-mini (cost: \$10, time: few hours)

## **The question:** What were people looking for?

- ▷ Conventional wisdom: "Craigslist was just for hookups"
- ▷ But was it? Let's look at the data.

This is an example of **text as data**—using NLP and LLMs to analyze written language. We'll learn how to do this ourselves on Tuesday.

## Classifying Intent: Romantic vs. Casual

We classified each post into mutually exclusive categories:

Section	Romantic	Casual	Platonic	Ambig.	R/C
Men seeking women	42%	20%	6%	32%	2.1
Women seeking men	48%	9%	6%	37%	5.7

**Finding:** Romantic posts outnumber casual for both genders.

The R/C ratio ( $\text{Romantic} \div \text{Casual}$ ) tells the story:

- ▷ Men: 2.1 romantic posts for every casual post
- ▷ Women: 5.7 romantic posts for every casual post

The conventional wisdom was wrong.

# The Gender Gap: A Theoretical Insight

Section	Romantic	Casual	Platonic	Ambig.	R/C
Men seeking women	42%	20%	6%	32%	2.1
Women seeking men	48%	9%	6%	37%	5.7
Gender gap in R/C					$5.7 - 2.1 = 3.6$

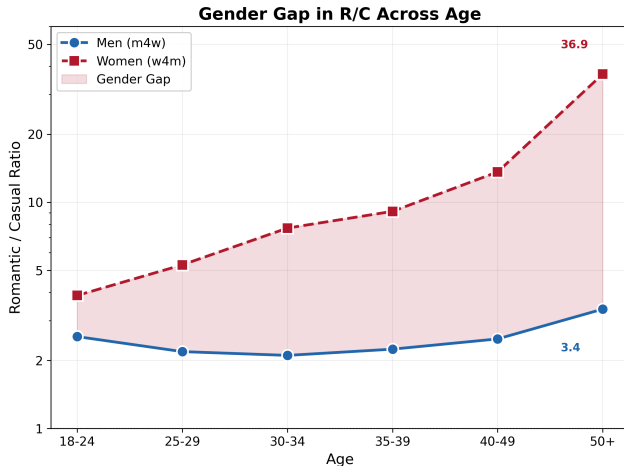
## Why this matters:

While  $R > C$  for both genders, the ratio is *much higher* for women.

- ▷ This creates a **shortage** of romantic men relative to romantic women
- ▷ More R-type women “seeking” R-type men than mathematically exist
- ▷ A gendered imbalance in the heterosexual online dating market

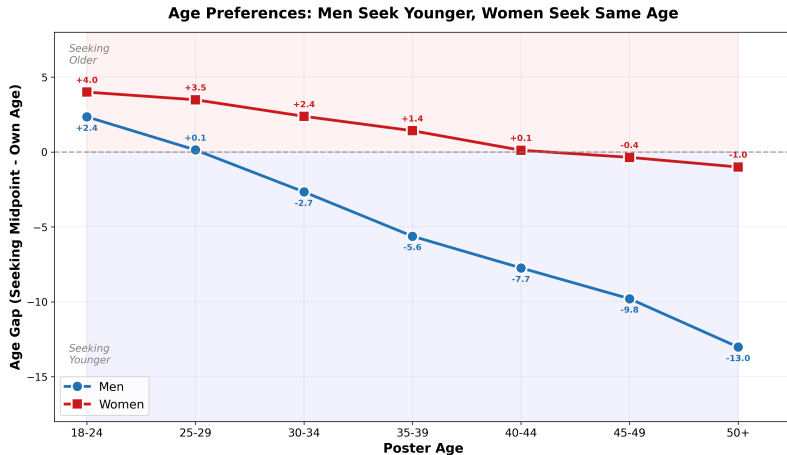
This had not been documented before—not in this way.

# The Same Story as a Picture



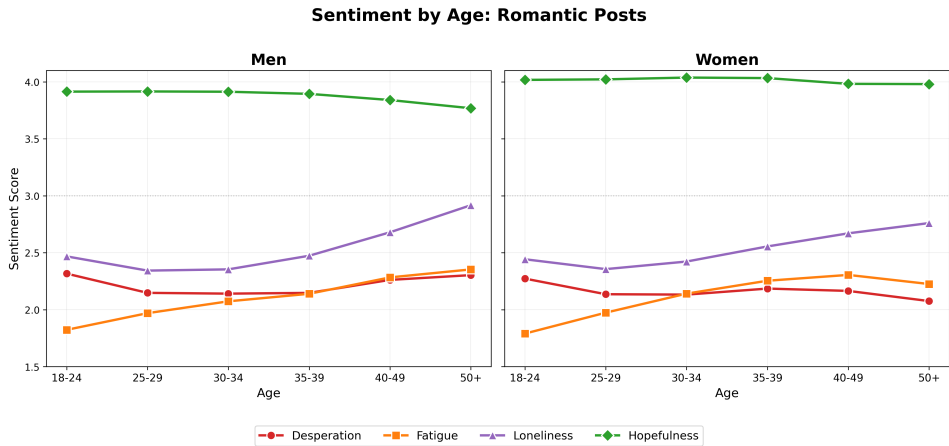
The shaded “gender gap” makes the divergence visceral.

# Age Preferences: A Diverging Pattern



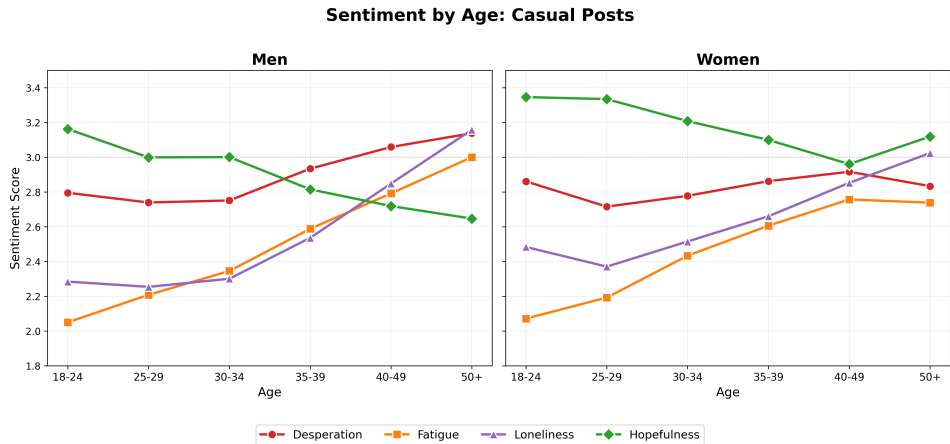
Men seek increasingly younger; women stay near their age. A table couldn't show this pattern.

# Sentiment by Age: Romantic Posts



I could put all 8 lines on one graph—but it would be unreadable. Facets keep it clear.

# Sentiment by Age: Casual Posts



Completely different pattern. The category (romantic vs. casual) changes the story.



# The Market Facing Older R-Type Women

Consider a **40-year-old woman** seeking a romantic relationship:

- 1. The pool of R-type men her age is small**
  - ▷ Men's  $R/C \approx 2.6$  at age 40–49; Women's  $R/C \approx 15$
- 2. R-type men her age prefer younger women**
  - ▷ Men 40–49 prefer women  $\approx 8$  years younger on average
- 3. C-type men targeting her are increasingly desperate**
  - ▷ Casual men's desperation rises sharply with age

**Her options:** Keep searching (delays childbearing) · Lower standards (unstable) · Give up

All three paths reduce fertility—  
even though  $R > C$  for both genders.

# What Made These Pictures Work?

1. **One idea per figure** — R/C ratio, age preferences, sentiment each get their own
2. **Facets for comparison** — Men vs. Women side by side, same scale
3. **Color with purpose** — Blue for men, red for women, consistent throughout
4. **Shading to emphasize** — The “gender gap” area draws the eye
5. **White space** — Clean, not cluttered
6. **Labels that explain** — Titles tell the story

Your figures should be readable without the text around them.



Numbers summarize. Visuals reveal. Use both.

# Looking Ahead

## Problem Set 1:

- ▷ Due Wednesday, February 11 at 11:59pm
- ▷ You now have all the statistics you need!
- ▷ Don't forget: GitHub URL in your document

## Next Week: Probability

- ▷ Foundation for everything that comes after
- ▷ Why we need it for inference

**Section this week:** Help with PS1, IPUMS setup