

Statistical Plots and Uncertainty

Sarah Groves

based on materials by Hannah Bako and Zhicheng Liu

Logistics

- Lab due **tonight**
- Sign up for **project presentation slot** – all next week
 - Part of your presentation grade is also asking questions of other presenters, so come ready to learn from your peers!
- Your final notebook does NOT have to be complete when you present; just highlight 1 or 2 findings

How today fits in

Question

What the viewer
wants to know
(TASK)



Plot

Data items → marks
Data attributes →
channels



Answer

What they walk
away with (GOAL)

Covered so far

- Bar chart/stacked bar chart
- Histogram
- Box plot
- Scatterplot
- Lineplot
- Heatmap
- Grouped plots

Today

How do we choose between different plot types for showing **distributions**?

How do we show statistical **relationships**?

When comparing single statistics (mean, median, etc.) how much **uncertainty** is in the answer?

*this is not a statistics class,
so we are only scratching the
surface on these topics!

What is your question really asking?

Interested in a shape?

“What do departure delays usually look like? Are they skewed?”



Distribution

histogram, density, box plot

Interested in a trend or relationship?

“Do countries with higher GDP have longer life expectancy?”



Relationship

scatter plot + fitted line

Comparing averages, medians, or some other *point estimate*?

“Do Gentoo penguins weigh more than Adélie penguins?”



Estimate + uncertainty

mean $\pm$ 95% CI, or show the distributions

Most questions start as a shape or a trend and turn into an estimate as soon as you compare groups.

What is your question really asking?

Distribution

histogram, density, box plot

- How closely do Elo-predicted win probabilities match actual game outcomes? (NBA, Parks)
- How does bike-sharing demand change across different seasons and months of the year? (Eric)

Relationship

scatter plot + fitted line

- How does the relationship between support availability and support satisfaction vary across the four domains (emotional, affectionate, tangible, positive social interaction)? (Renisha)
- Does draft position still affect salary later in a career? (Nahom)

Estimate + uncertainty

mean $\pm$ 95% CI, or show the distributions

- Did the post 2020 recovery in international arrivals match the recovery in total tourism spending, or did tourists spend differently after the pandemic? (Clairine) (Comparing distributions)
- Did home-field advantage hold up across the season, and did it change month to month? (Samuel) (Average + uncertainty)

Most questions start as a shape or a trend and turn into an estimate as soon as you compare groups.

Pick two of your own questions

3 min

Q1 · a shape or trend question

My question: _____

Shape or trend? _____

Variable(s): _____

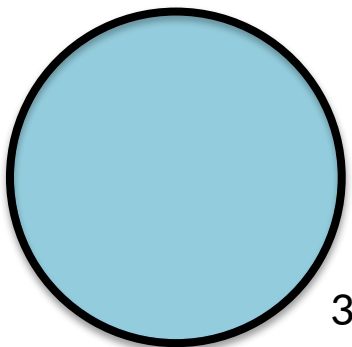
Q2 · an estimate that has uncertainty

My question: _____

What am I estimating? (a mean, a difference, a rate)

Groups being compared: _____

Take them from your Project 1 Checkpoint #2 or new ones that have come up as you've started working on your Project 1 EDA. We'll come back to Q1 and Q2 throughout today.

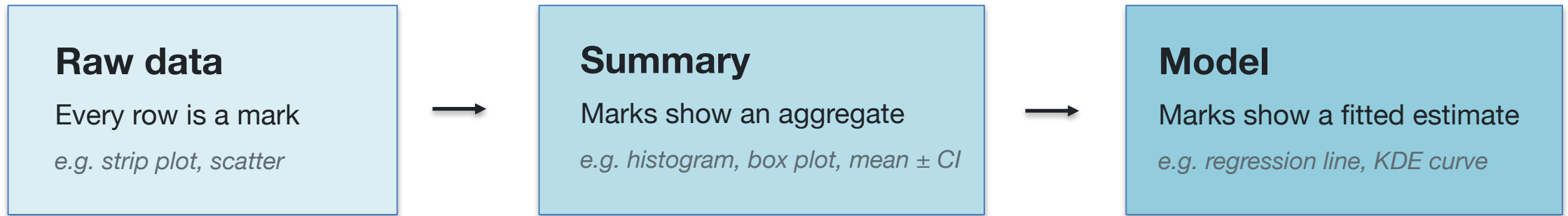


3 min timer

Statistical plots

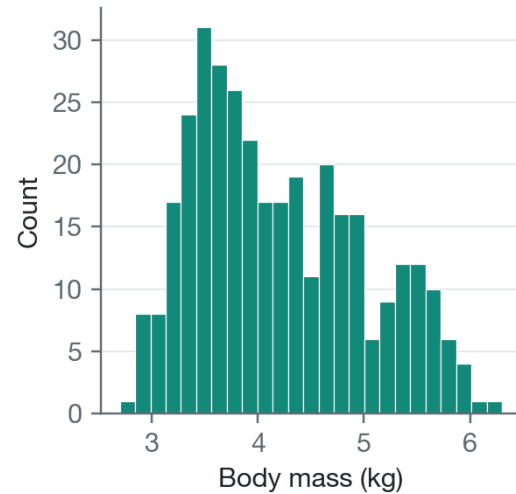
When the answer is a **shape** or a **trend**, not a single value

What makes a plot “statistical”?

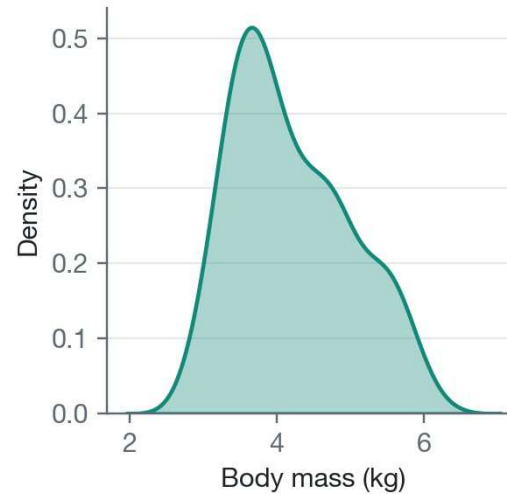


Each step to the right hides more of the raw data and **adds assumptions**: bin width, smoothing bandwidth, a straight-line model.

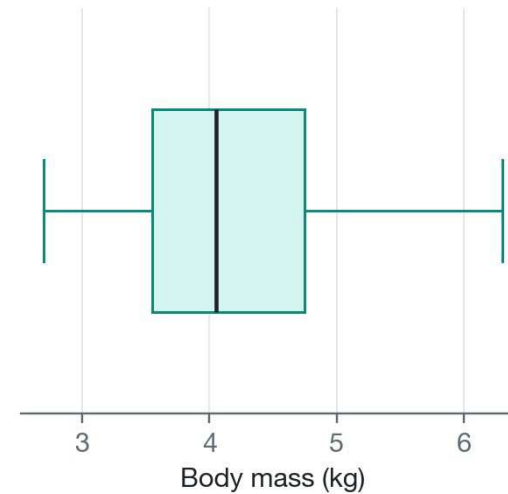
Questions about one variable: show the **distribution**



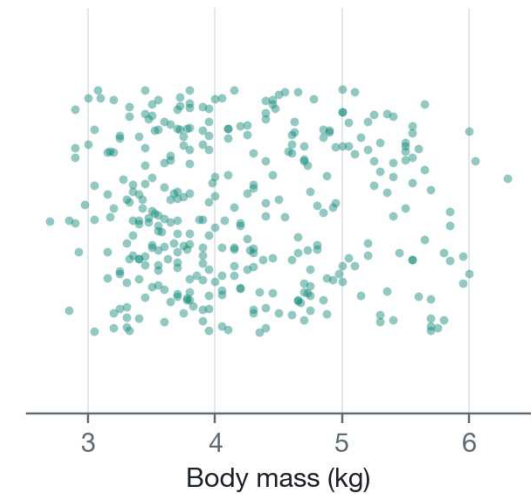
Histogram
counts in bins



Density (KDE)
smoothed shape



Box plot
median, IQR, outliers



Strip / swarm
every point

Good at... Shape, skew, gaps

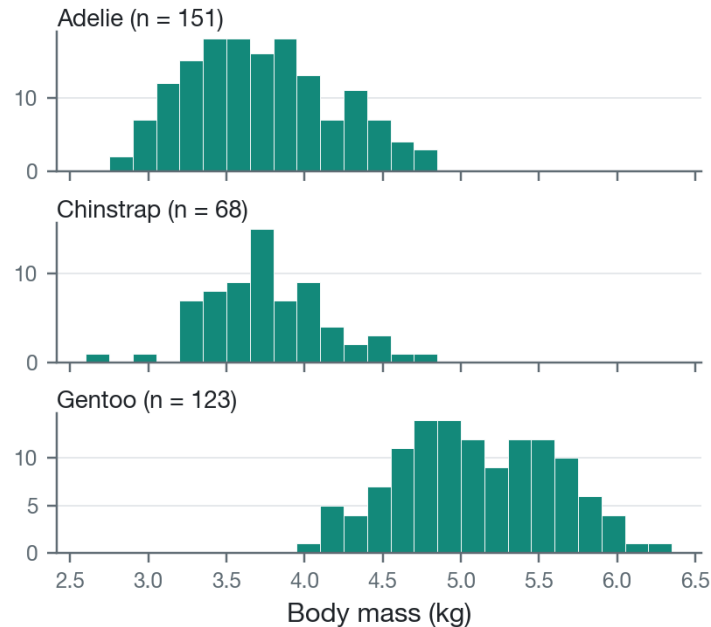
Smooth shape;
overlaying groups

Comparing many
groups at once

Honest about
number of points

Use when the question is: What is typical? How spread out is it? Is it skewed, or are there multiple humps?
Any outliers?

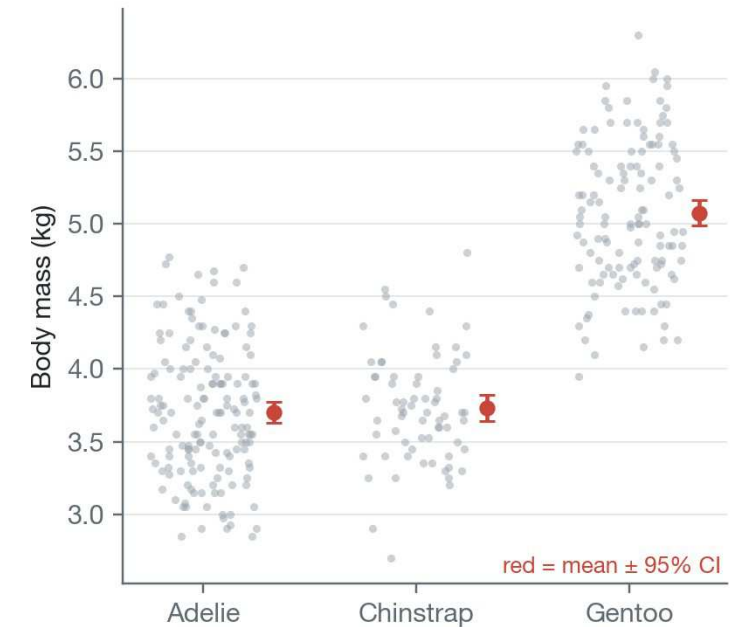
Comparing distributions across groups



Best for seeing each shape



Violin: shape + comparison



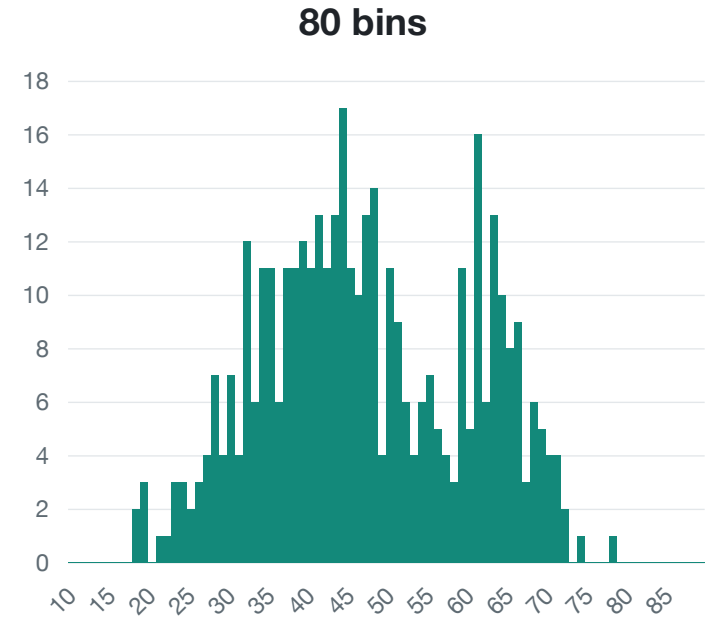
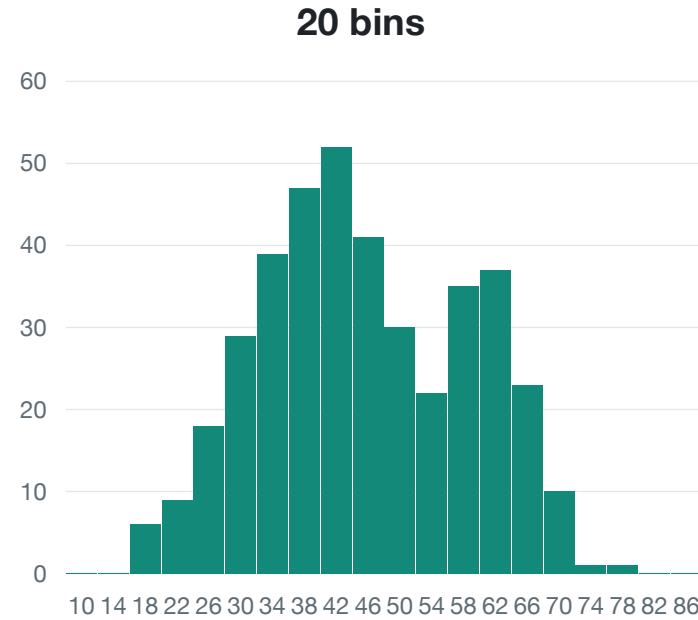
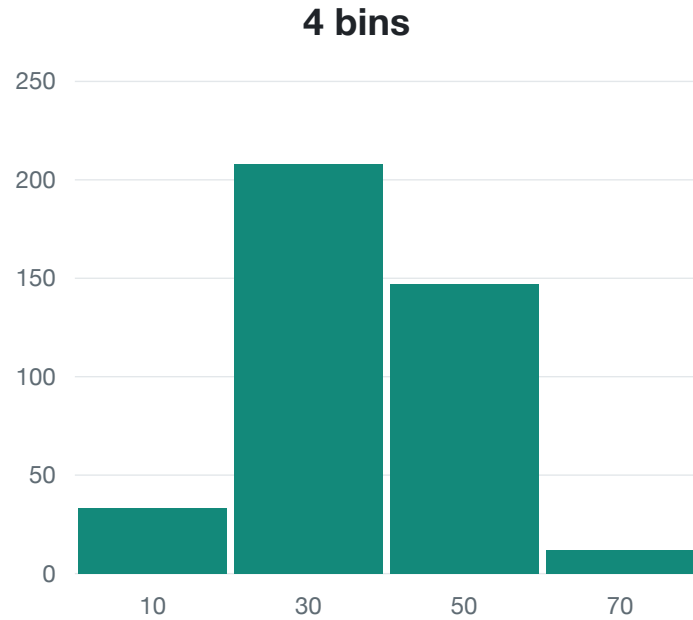
Best when n is small

Ask: is the difference in the center, the spread, or the shape? Align the axes so the reader can compare.

Choosing a distribution plot

Plot	Good at	Hides	Choice you have to make
Histogram	Shape, skew, gaps	Detail within a bin	Bin width / number of bins
Density (KDE)	Smooth shape; overlaying groups	Sample size; can spill past real limits	Bandwidth
Box plot	Comparing many groups at once	Bimodality, n per group	Whisker rule ($1.5 \times \text{IQR}$)
Violin	Shape + comparison	n; looks precise even when n is small	Bandwidth, scaling
Strip / swarm	Honest about n	Shape when n is large	Jitter, transparency

Same data, three bin widths



Try several bin widths before you pick one

Back to your Q1: if it's a shape

1–2 min

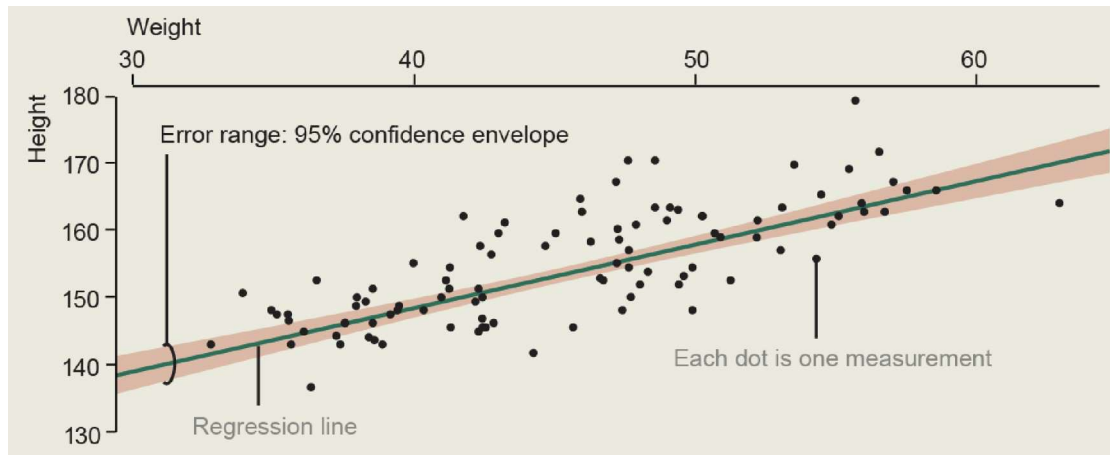
Q1 · your shape or trend question

- 1 Which distribution plot fits best: histogram, density, box plot, or strip plot?
- 2 What's the setting you'll need to test (bin width, bandwidth)?
- 3 What would the plot hide? (e.g., a box plot hides two humps in the distribution)

You can jot your answers in your Project 1 notebook or on paper.

Example: “Departure delays” → histogram; try 20, 50, 100 bins; check whether a log scale shows the long tail.

Questions about two variables: show the **relationship**



Use when the question is:

Does Y go up when X goes up? How strong is the link? Does it differ by group?

Design choices

- Overplotting: use transparency, smaller marks, or a hexbin (shows where the points fall in bins, like a 2D histogram)
- Trend: show a linear relationship with regplot; more curvy best fit lines are also possible
- Groups: use hue or small multiples (`sns.regplot(hue=...)`)
- Scale: log axes when values span orders of magnitude

Back to your Q1: if it's a trend

1–2 min

Q1 · your shape or trend question

- 1 What goes on x and on y? Does either need a log scale?
- 2 Will points overlap? (transparency, hexbin, sampling)

Jot your answers in your Project 1 notebook or on a piece of paper.

Example: “GDP vs life expectancy” → scatter, log x, color by continent; say “associated with,” not “causes.”

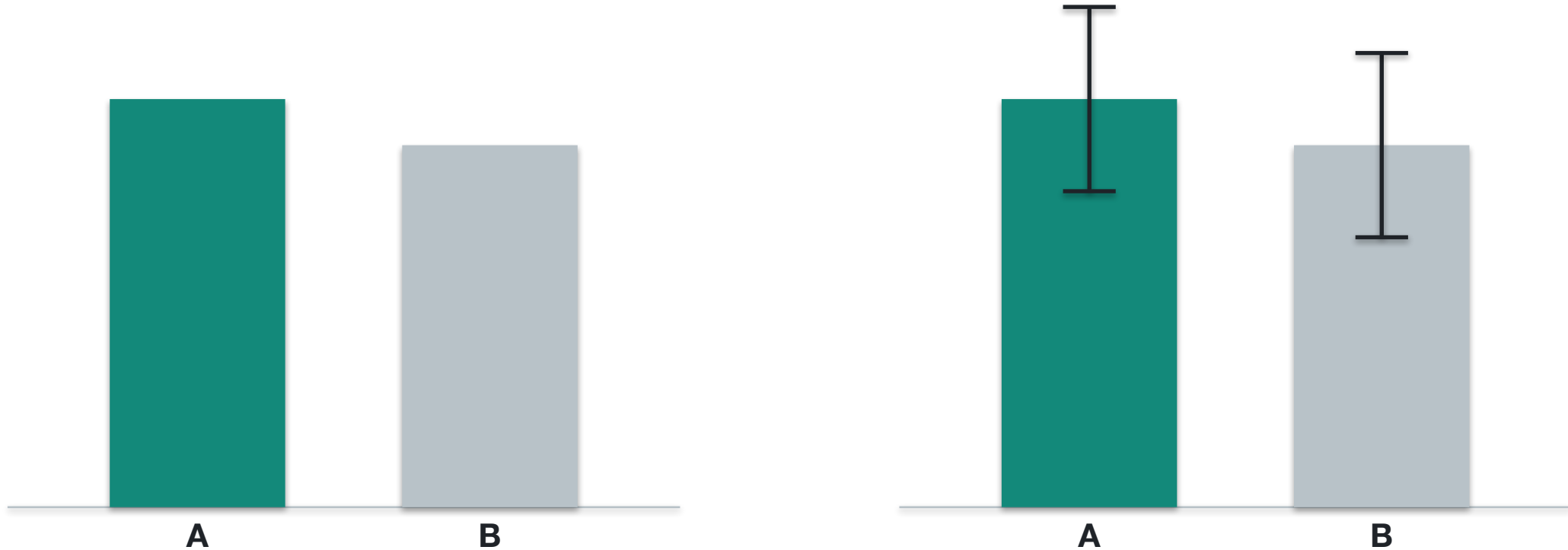
Visualizing uncertainty

Every summary is an estimate. How sure are we?

This bar chart is showing the mean of group A vs the mean of group B. Is A bigger than B?

Looks like a clear yes

With 95% CIs (n = 12 per group)



A point estimate with no interval looks more certain than the data can support.

Where does uncertainty come from?

Acquisition

- 1 Sampling (surveys, exit polls) and measurement error, from people and from instruments. Who is missing? More measurements, more confidence.

Transformation

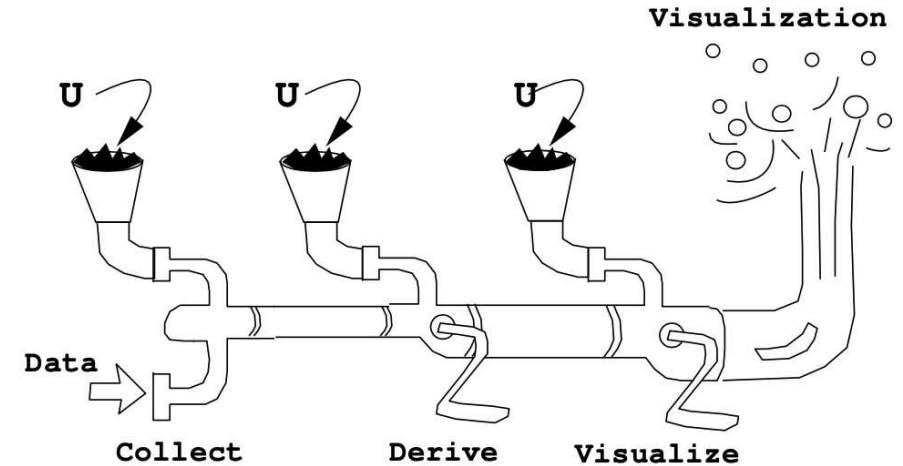
- 2 Unit conversion, aggregation, rescaling, binning, filtering. A model is a simplification of the data.

Visualization

- 3 Interpolated lines imply values nobody measured. Color and size aren't perceived linearly.

Forecast

- 4 Predictions: quantitative (next bus arrival) or categorical (rain / sun / cloud).



Uncertainty (U) enters at every stage of the pipeline, not just at collection.

Back to your Q2: where's the uncertainty?

1–2 min

Q2 · your estimate that needs uncertainty

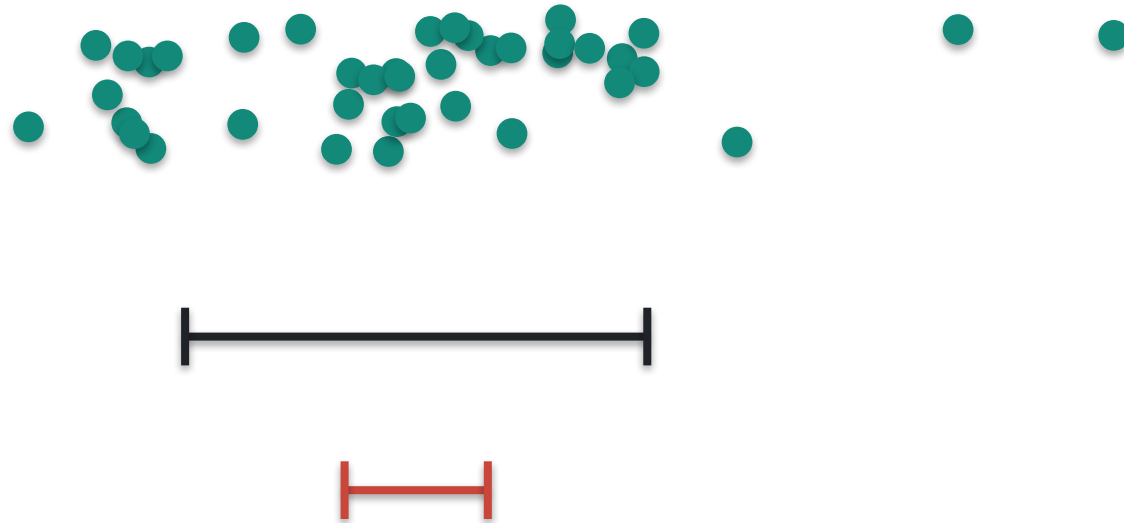
- 1 Is there uncertainty coming from data acquisition, aggregation into groups, binning, or filtering?
- 2 Are values missing? Is anything missing from the data entirely?
- 3 Are any of your groups in your data very small?

Jot your answers in your Project 1 notebook or on a piece of paper.

Example: “Gentoo vs Adélie body mass” → $n = 124$ vs 152 ; 2 missing masses dropped.

Variability $\neq$ uncertainty

40 observations

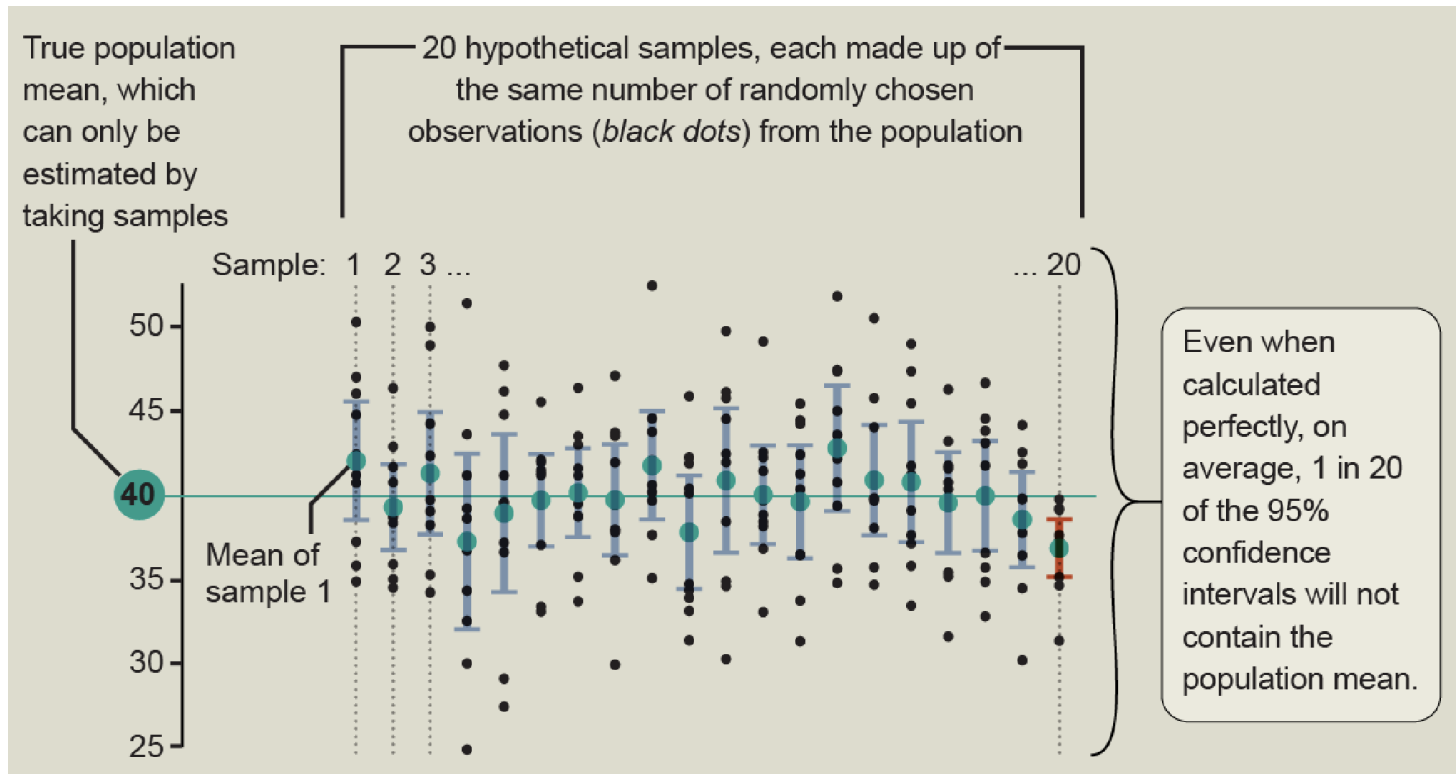


± 1 SD
spread of the data

95% CI of the mean
uncertainty in the estimate

More data shrinks the CI, but not the SD. Say which one your error bars show in a label or caption!

What does a 95% CI actually mean?



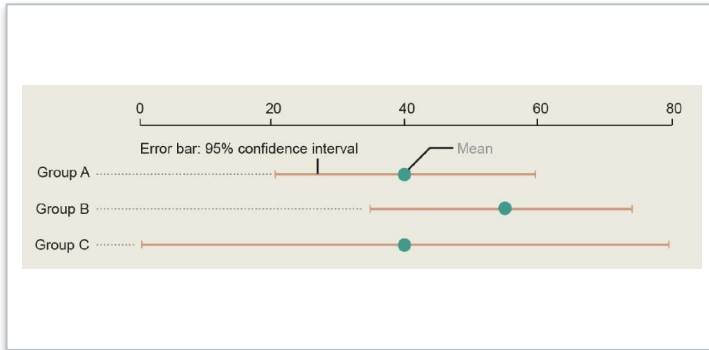
Common reading

“There's a 95% chance the true value is in this interval.”

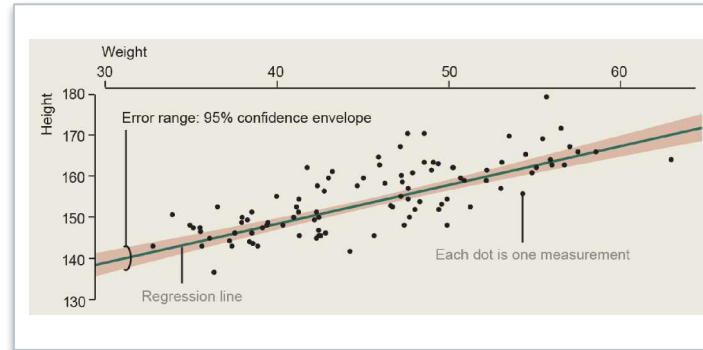
Closer to right

If we repeated the sampling many times, about 95% of the intervals would contain the true value.

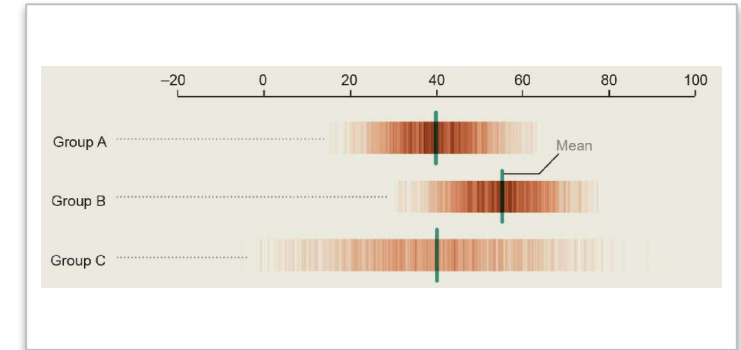
A toolkit for showing uncertainty



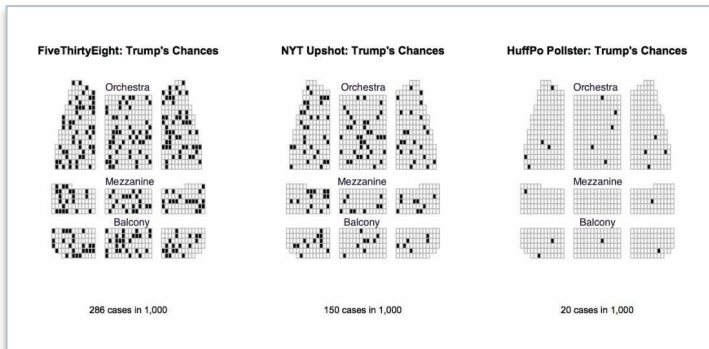
Error bars
mean $\pm$ CI / SE / SD



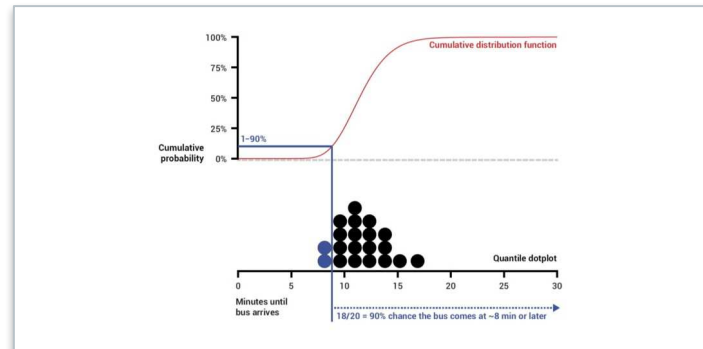
Confidence band
around a fitted line



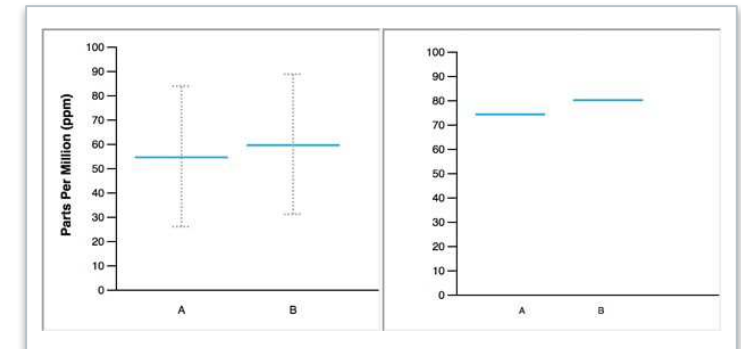
Gradient plot
darker = more likely



Icon array
“286 cases in 1,000”



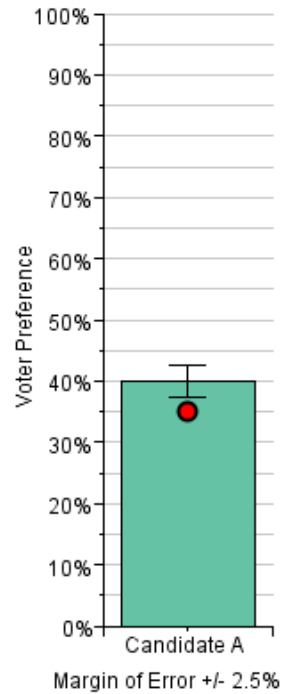
Quantile dot plot
each dot = 1 in 20 outcomes



Hypothetical outcomes (HOPs)
animated draws

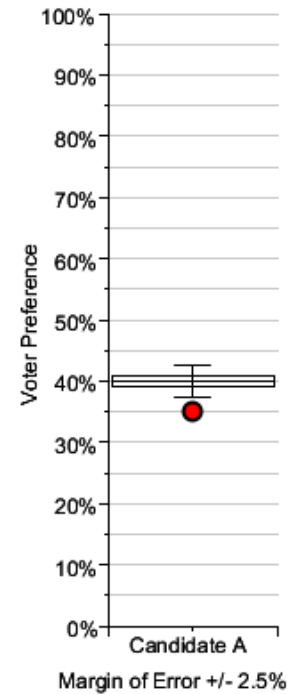
Task: How likely is the red outcome?

Poll: Candidate A at 40%, margin of error $\pm 2.5\%$. Is the red dot a typical outcome or an unusual one? Rate 1 (very unlikely) to 7 (very likely).



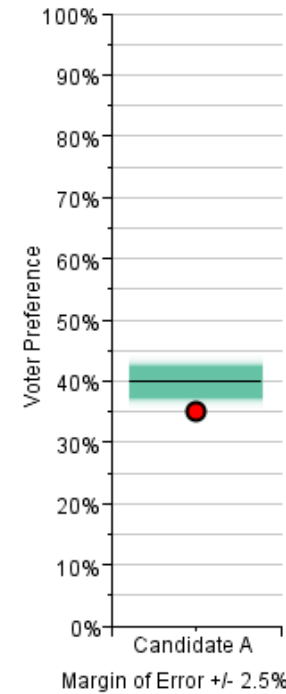
1 Bar + error bar

1 2 3 4 5 6 7



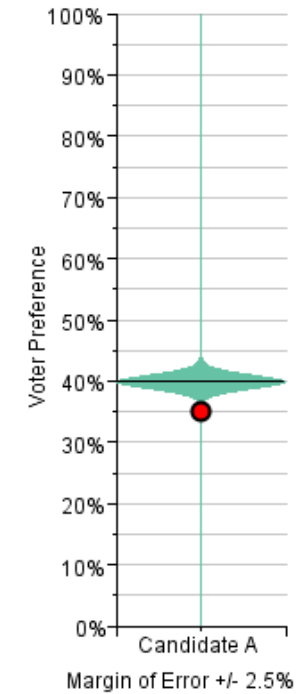
2 Box plot

1 2 3 4 5 6 7



3 Gradient

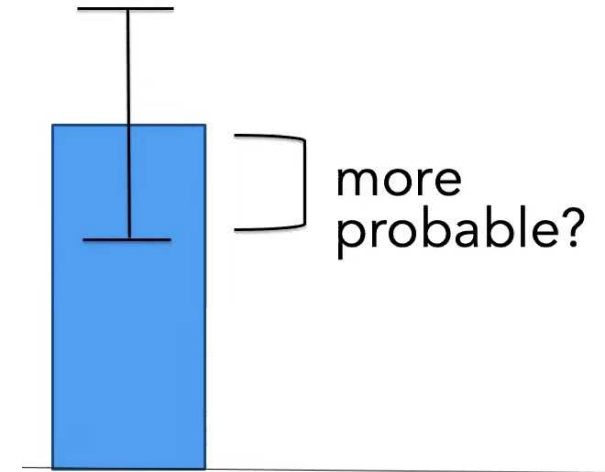
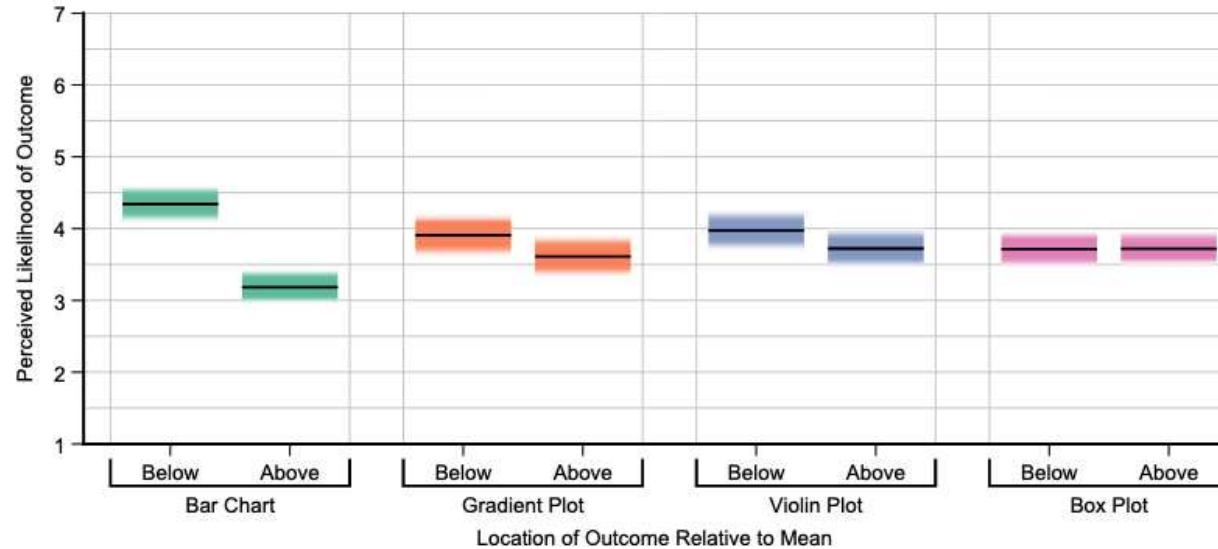
1 2 3 4 5 6 7



4 Violin

1 2 3 4 5 6 7

Results: the bar changes what feels likely



Outcomes above and below the mean should feel equally likely.

With bar charts they don't: points inside the bar are rated likelier than points outside it.

Back to your Q2: sketch it two ways

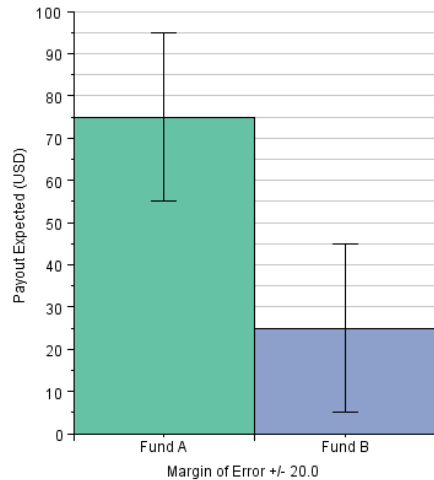
1–2 min

Q2 · your estimate that needs uncertainty

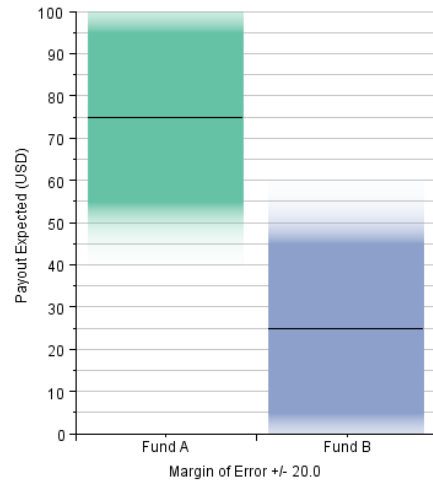
- 1** Sketch Q2 as a bar chart with error bars.
- 2** Sketch it again as points + mean $\pm$ CI, or a violin / gradient.

Jot your answers in your Project 1 notebook or on paper.

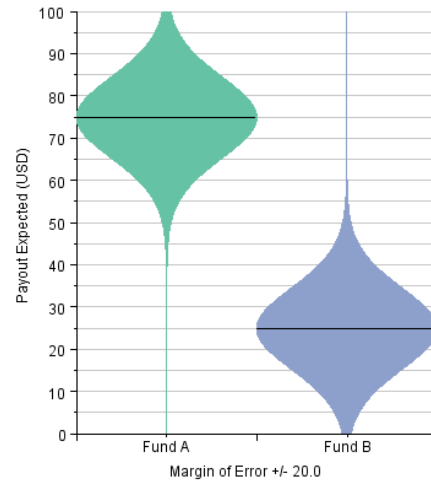
Error bars: read the fine print



Bar + error bar



Gradient



Violin

Rules of thumb

- Label what the bars are: SD, SE, or 95% CI
- Report n per group
- Present intervals as a range of plausible values, not a hard edge
- Overlapping 95% CIs $\neq$ “no difference”
- Show the points when n is small

Within-the-bar bias: values inside the bar look likelier than values above it.

Binary reading: a value is either “in” or “out”, and effect sizes look bigger than they are.

2016: what's your impression?

Predictions from 2016 presidential election

[Justin H. Gross, Washington Post, <http://wapo.st/2fCYvDW>]

FiveThirtyEight

28%

NYT Upshot

15%

HuffPo Pollster

2%

Three forecasts of Trump's chance of winning. If you'd seen these before the election, what would you have expected to happen?

People read counts better than percentages

Predictions from 2016 presidential election

[Justin H. Gross, Washington Post, <http://wapo.st/2fCYvDW>]

FiveThirtyEight

NYT Upshot

HuffPo Pollster

28%

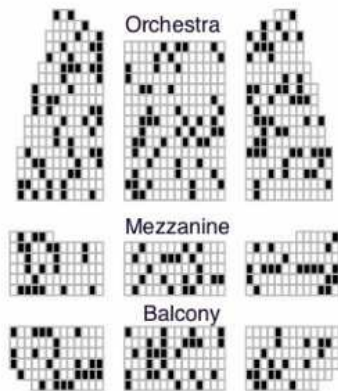
15%

2%

Why this works

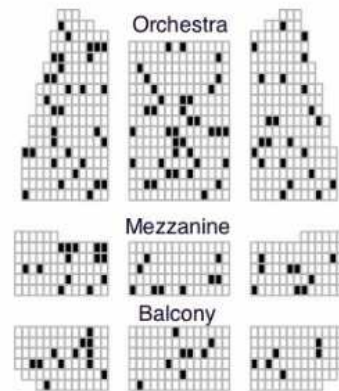
- “2%” and “15%” both feel like “won't happen”
- Icon arrays turn a probability into a count you can see
- People judge area poorly, so counts beat circles and densities
- One icon array shows one probability

FiveThirtyEight: Trump's Chances



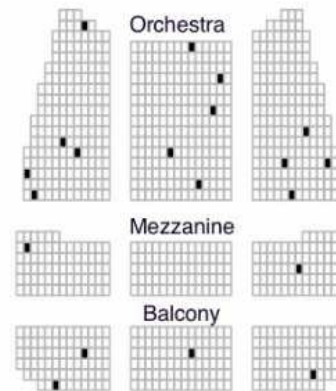
286 cases in 1,000

NYT Upshot: Trump's Chances



150 cases in 1,000

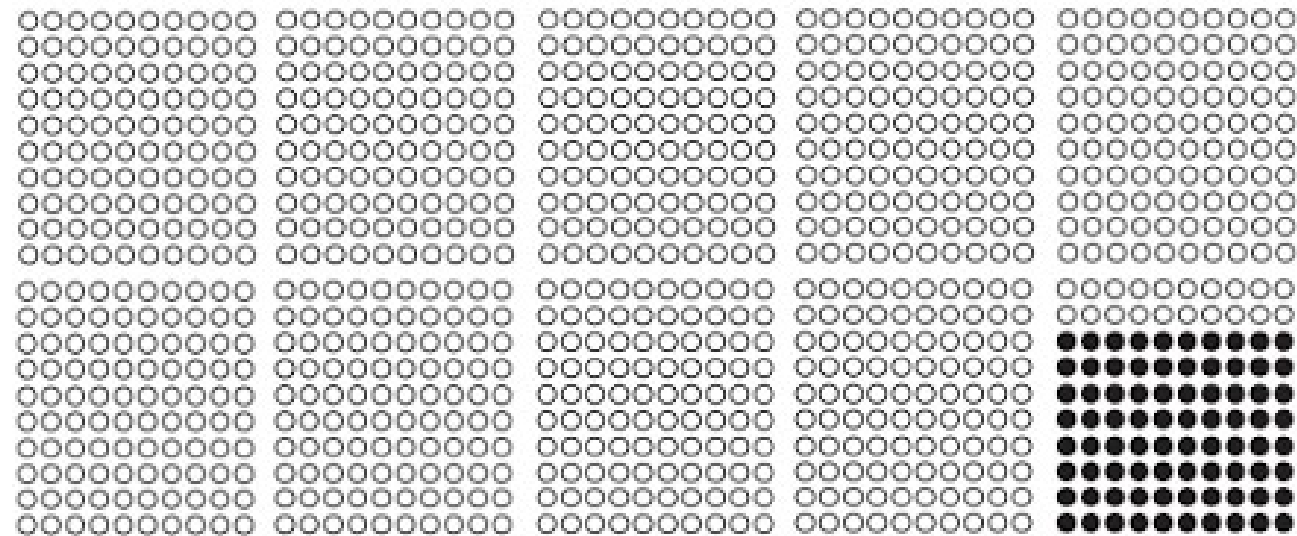
HuffPo Pollster: Trump's Chances



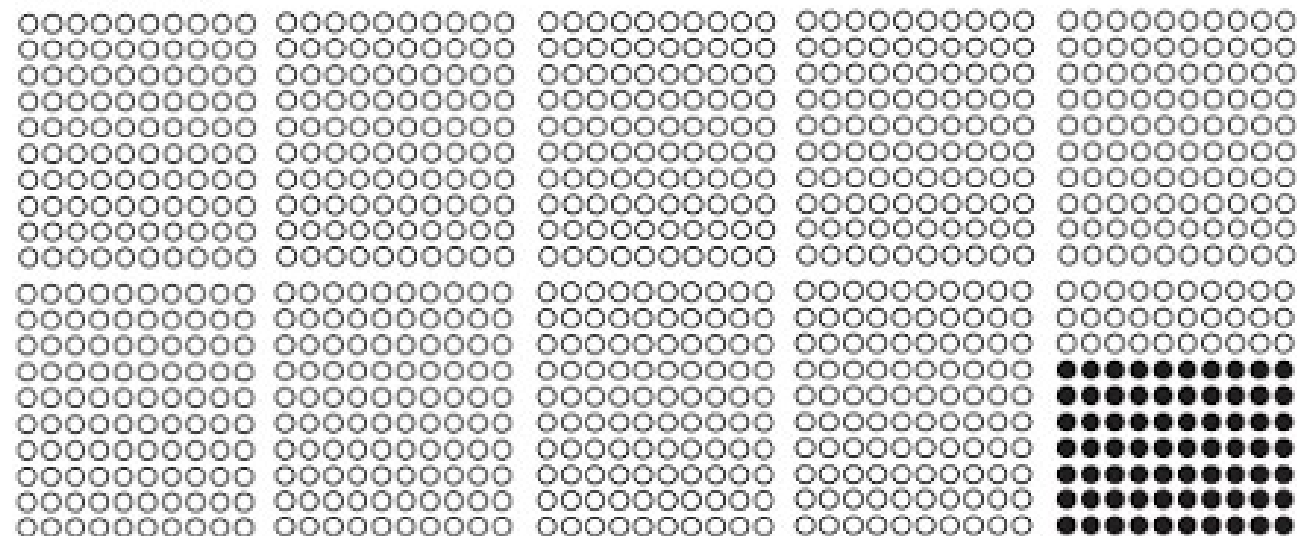
20 cases in 1,000

For people with symptoms of arterial disease, aspirin can reduce the risk of having a stroke or heart attack by 13%.

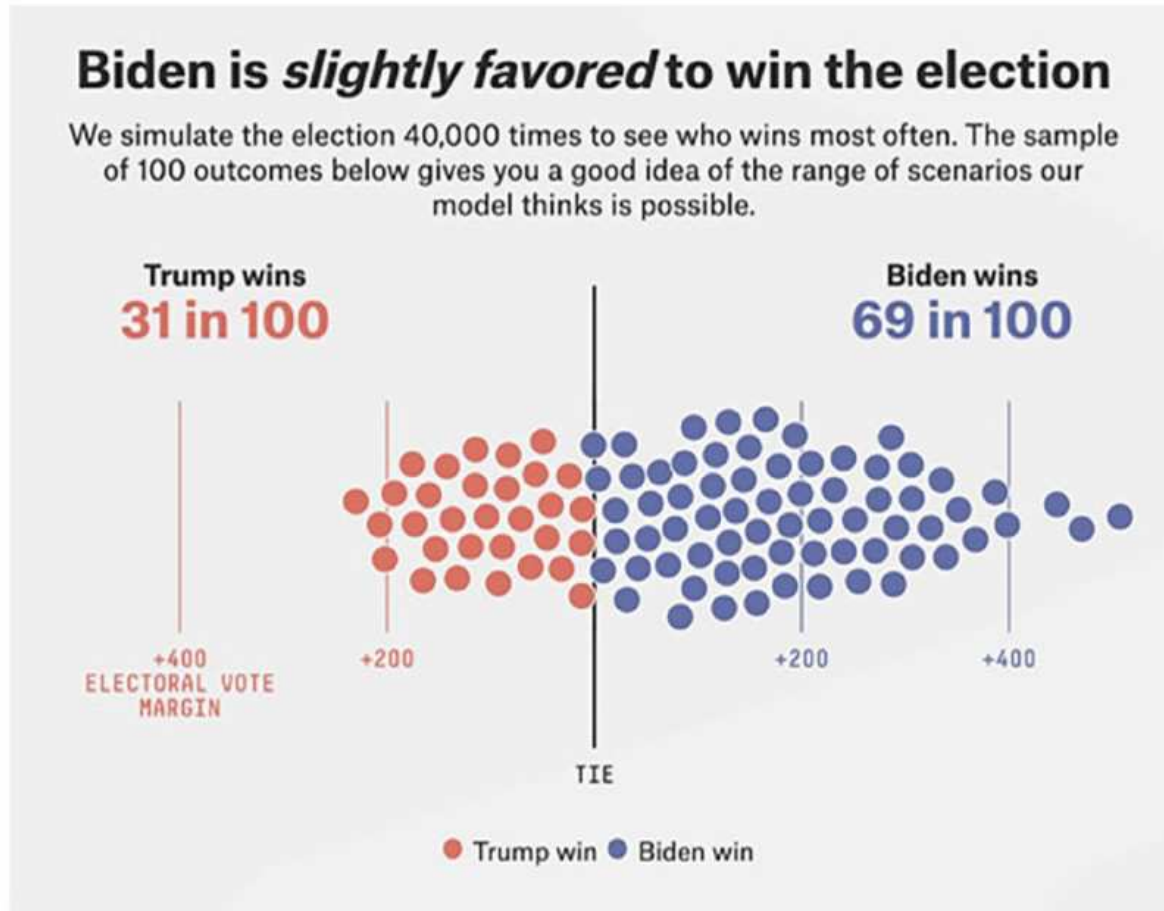
Without aspirin



With aspirin



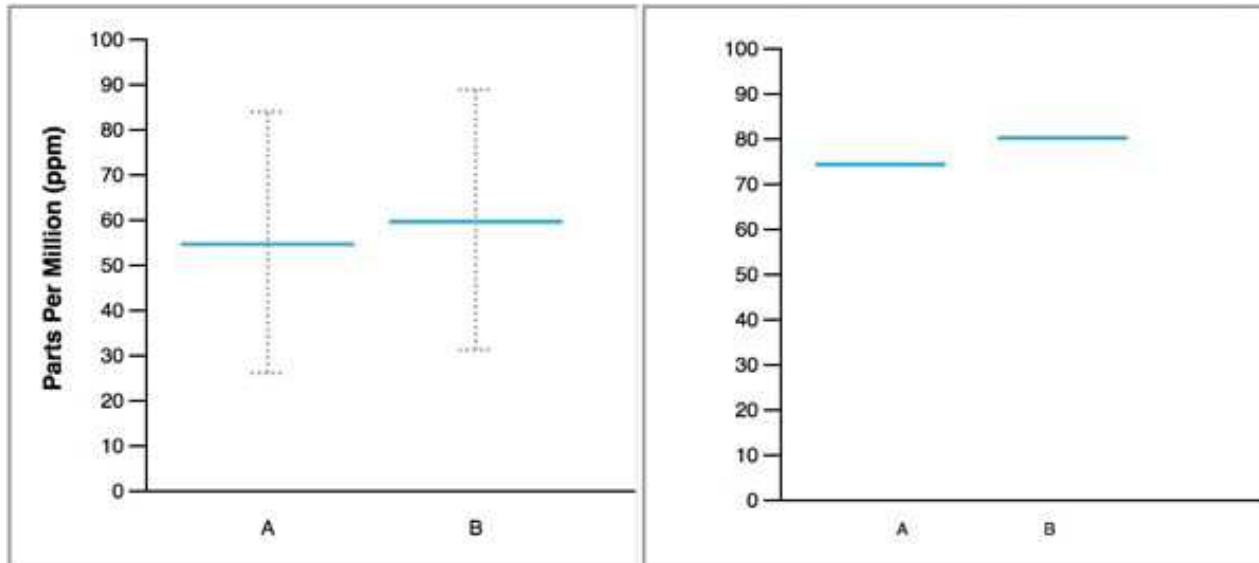
In the wild: 538's 2020 election forecast



Instead of one big percentage: 100 simulated outcomes, one dot each (a “ball swarm”).

Odds in words (“1 in 10”) rather than decimals.

Animation: HOPs and jittery gauges



Watch the right panel: *how often is B bigger than A?* HOPs show one possible outcome per frame.



Jittery gauge (NYT, 2016)

- Good for live prediction
- Uncertainty can't be ignored
- Provoked anxiety

In the wild: the monthly jobs report

If the economy actually added 150,000 jobs last month, it would be possible to see any of these headlines:

The jobs number is just an estimate, and it comes with uncertainty.

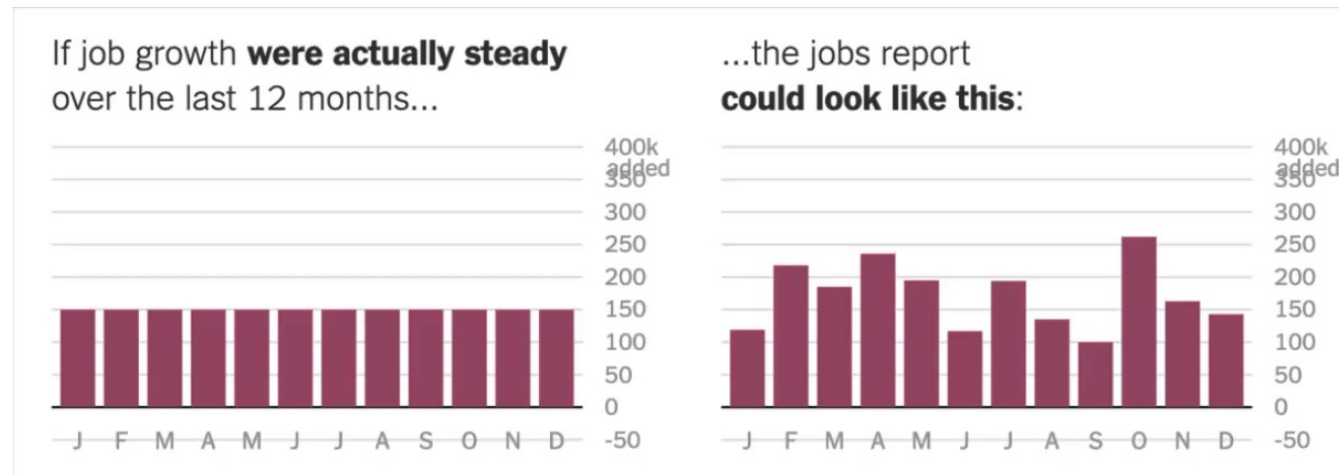


The BLS's 90% interval on one month's job change is roughly $\pm 100,000$ jobs.

The NYT simulated reports from an economy with steady growth: some months look like a boom, others like a slump.

In Aug 2024, BLS revised March 2024 employment down by 818,000 jobs.

The report says +150,000 jobs. Is the economy growing?



Source: [NYT Upshot \(2014\)](#) · [BLS](#) · [2024 revision](#)

Say it in the caption, too

Overclaims

“Gentoo penguins are heavier than Adélie penguins.”

“Delays increase with distance.”

Calibrated

“In this sample ($n = 342$), Gentoo penguins averaged $\sim 1,400$ g more (95% CI: [], []).”

“Delays are weakly associated with distance ($r \approx []$).”

Caption checklist

- What the interval or band means (CI / SE / SD)
- Sample size, and anything missing or excluded
- “Associated with”, not “causes”
- Frequency framing for a general audience: “about 1 in 5”

How to do it in Python

```
# mean + 95% CI by group (seaborn bootstraps the CI)
sns.pointplot(data=df, x="species", y="body_mass_g",
              errorbar=("ci", 95))
sns.stripplot(data=df, x="species", y="body_mass_g", alpha=0.3)

# scatter + regression line + CI band
sns.regplot(data=df, x="flipper_length_mm", y="body_mass_g")

# matplotlib: error bars and bands by hand
ax.errorbar(x, means, yerr=1.96 * ses, fmt="o", capsize=4)
ax.fill_between(x, lower, upper, alpha=0.2)

# bar plot of maximums with CI band
sns.barplot(data=df, x="category", y="measurement",
            estimator=np.max, errorbar=("ci", 95))
```

seaborn errorbar=

- ("ci", 95): uncertainty in the mean
- "se": standard error
- "sd": spread of the data
- ("pi", 50): middle 50% of the data
- None: turn it off

Match the plot to the **question**
Don't hide the **distribution** if it helps
answer your question
Then show how **certain** you are about the
answer
Avoid **overclaiming**