Welcome to STA 101!

Statistics is a confrontation with uncertainty.

Statistics confronts uncertainty by quantifying it.

Data analysis

Transforming messy, incomplete, imperfect data into knowledge.

Statistical inference

Quantifying uncertainty about that knowledge.

Do what you can do, at the pace you can do it

		Technical facility	
		No	Yes
Conceptual	No	Oh well!	Dangerous!
Understanding	Yes	Not bad!	Great!

Updates

- Office hours;
- Ed Discussion;
- **Request materials!**

Data

- Data is any information you collect about the world;
- Often represented in a spreadsheet (AKA `data frame`):

	Subject	Variable 1	Variable 2	Variable 3
1	1	0.06309445	-0.51459791	0.7205242
2	2	0.35845812	0.03989746	0.6137163
3	3	-0.24366289	0.41237579	1.5903795
4	4	0.15558421	0.50109301	0.2527414
5	5	0.87404174	1.71982009	0.4403238

- rows = **observations**; columns = **variables**;
- number of rows = **sample size**.

(JZ commentary: "Big data" is when this thing is so massive that you could not even store the file on a single computer.)

Example: political poll

person	state	sex	college	vote
1	TX	M	TRUE	idk
2	SC	M	TRUE	Harris
3	TX	F	TRUE	idk
4	PA	M	FALSE	Harris
5	NY	M	FALSE	Trump
6	NC	F	TRUE	Trump
7	OH	F	FALSE	Harris
8	TX	M	TRUE	Trump

Observational unit = person

Example: Flint, MI water data

	id	zip	ward	lead	draw
1	1	48504	6	0.344	first
2	1	48504	6	0.226	second
3	1	48504	6	0.145	third
4	2	48507	9	8.133	first
5	2	48507	9	10.770	second
6	2	48507	9	2.761	third
7	4	48504	1	1.111	first
8	4	48504	1	0.110	second
9	4	48504	1	0.123	third
10	5	48507	8	8.007	first
11	5	48507	8	7.446	second

Observational unit = household/draw

(JZ commentary: asking "what is the observational unit in this data set" or "what does a row correspond to" is never a bad question!)

Data analysis

- Transforming incomplete, imperfect data into knowledge;
- This messy, trial-and-error process often resembles **art** and **rhetoric** more than science;
- What form does the knowledge take?
    - pictures;
    - a concise set of numerical summaries.
- **ABV**: **A**lways **B**e **V**isualizing.

(JZ commentary: We don't have "big data" in this class, but even "small data" is too much for the human mind to comprehend. We need to compress the data down into pictures and numerical summaries to appreciate the information. And you need both, working in tandem. If you just look at pictures, someone will eventually tap you on the shoulder and ask "great, but how much insulin are we actually giving the patient?" If you just look at numerical summaries, you can be mislead: see Anscombe's quartet later.)

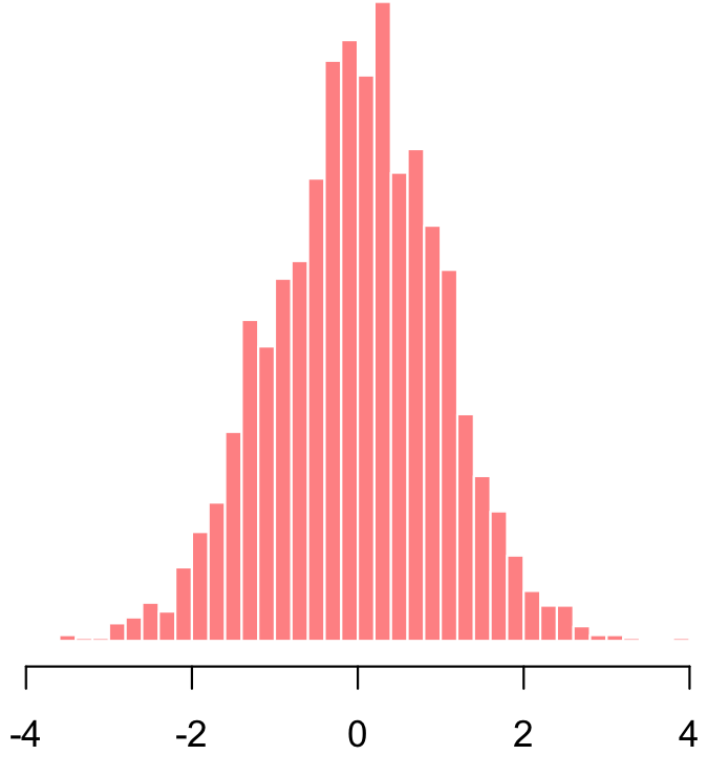
Baby's first data visualization: the histogram

Histogram: displays the **distribution** or **variability** of the data. Where are the values typically concentrated? How spread out are they? Are there asymmetries?

Turn this...

Subject	Variable 1
1	0.74433805
2	-0.95860459
3	-1.05801249
4	0.50451677
5	-0.88977189
6	-0.35737586
7	0.31391182
8	0.81459953
9	0.24745903

...into this

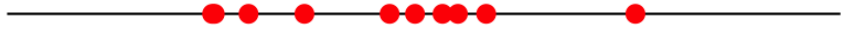


How is a histogram drawn?

1. Start with the number line (horizontal axis);
2.
3.
4.
5.

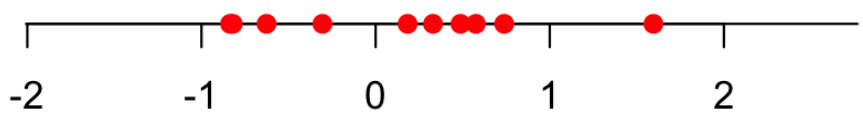
How is a histogram drawn?

1. Start with the number line (horizontal axis);
2. Put your data values on the line;
3.
4.
5.



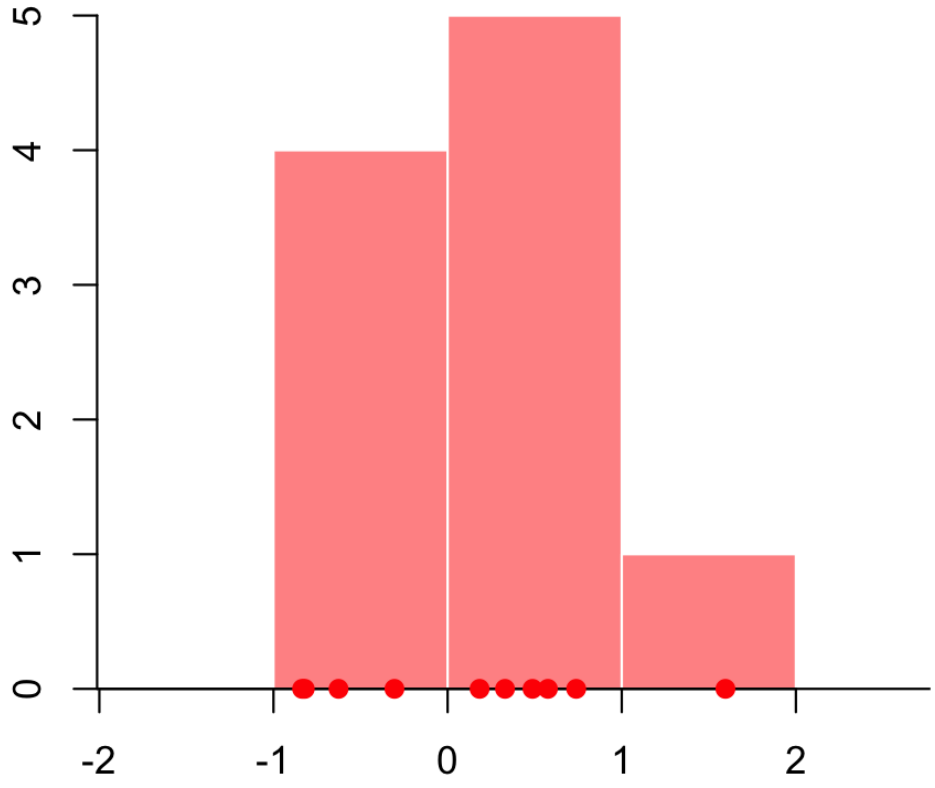
How is a histogram drawn?

1. Start with the number line (horizontal axis);
2. Put your data values on the line;
3. Break the line into bins;
4.
5.



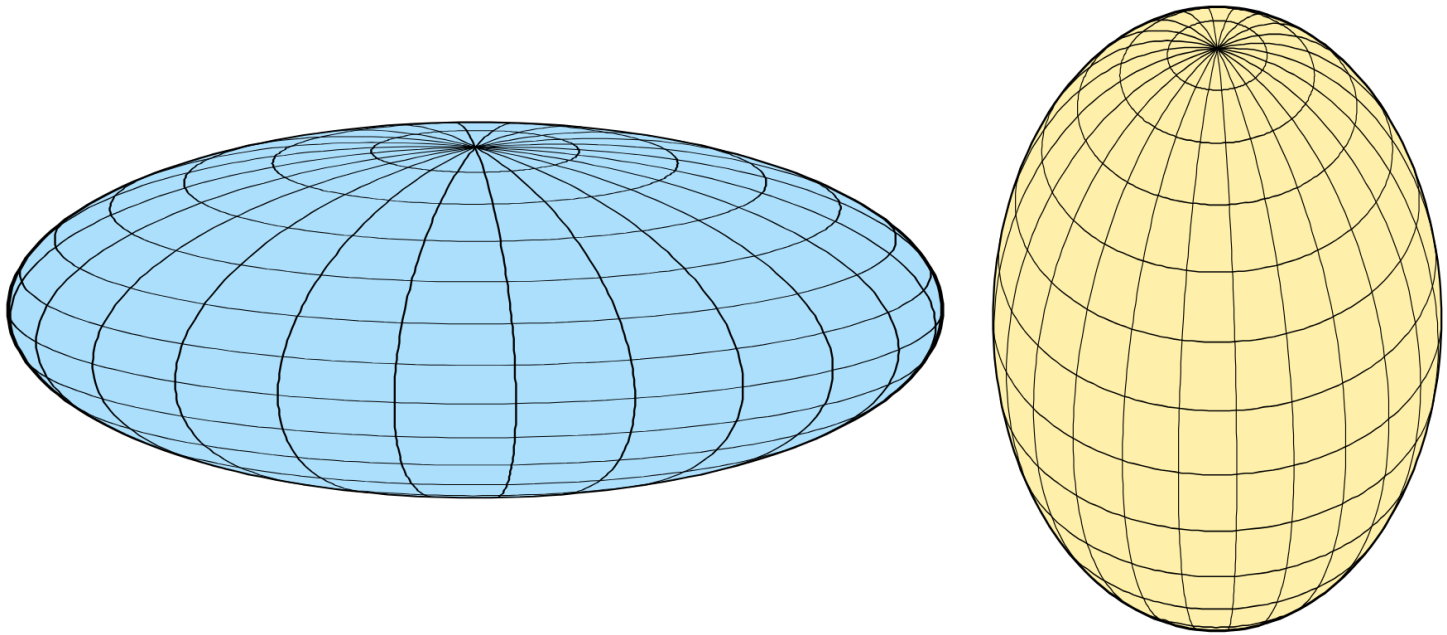
How is a histogram drawn?

1. Start with the number line (horizontal axis);
2. Put your data values on the line;
3. Break the line into bins;
4. Count how many data values fall in each bin;
5. Put bars over the bins (height = count).



Some history: measuring the shape of the Earth

Is the Earth perfectly spherical, or is it smushed? (it's smushed)



(JZ commentary: see Stigler's *Seven Pillars of Statistical Wisdom* for details of this example.)

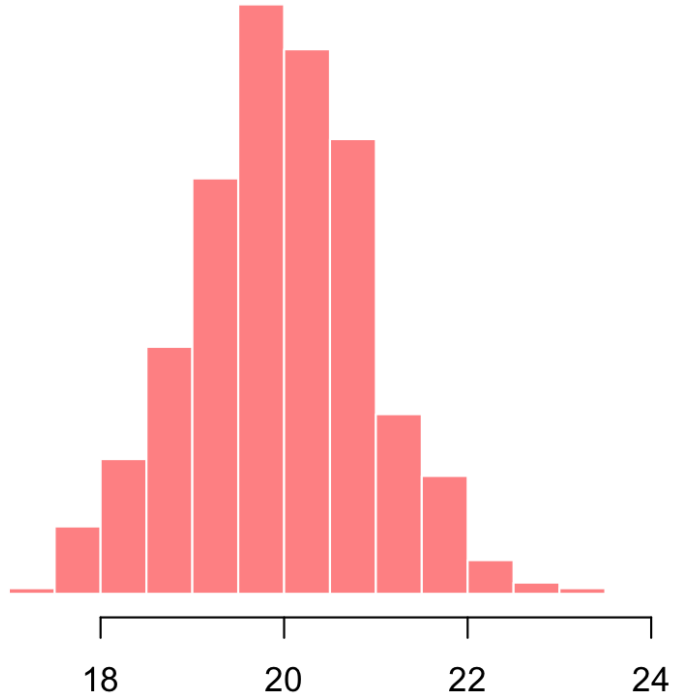
Earth: smushed? (yes)

You can determine this by watching a heavenly body move through the night sky and measuring the length of the arc that it traces:

Several research teams across the world did this in the 1700s and came up with one of the first spreadsheets:

Locus obſervationis	*Latitudo* °	′	*½ ſin. vers. rad.* 10000	*Hexapedæ*	*Differ. a primo*	*Differ. computata*	*Error*
In America	0	0	0	56751	0	0	0
Ad Prom. B. S.	33	18	2987	57037	286	240	−46
In Italia	42	59	4648	56979	228	372	144
In Gallia	49	23	5762	57074	323	461	138
In Lapponia	66	19	8386	57422	671	671	0

The data (I made this up)



- Why didn't they all get the same number? (human error, imprecision in the measurement instruments, etc);
- Given this, what's our best guess? ("the middle")

(JZ commentary: This is not the answer they would have given back then. They would instead decide which row in the spreadsheet is "best" and discard the rest. Times have changed!)

Summary statistics

Measures of "center":

- **mean**: average of the data values...

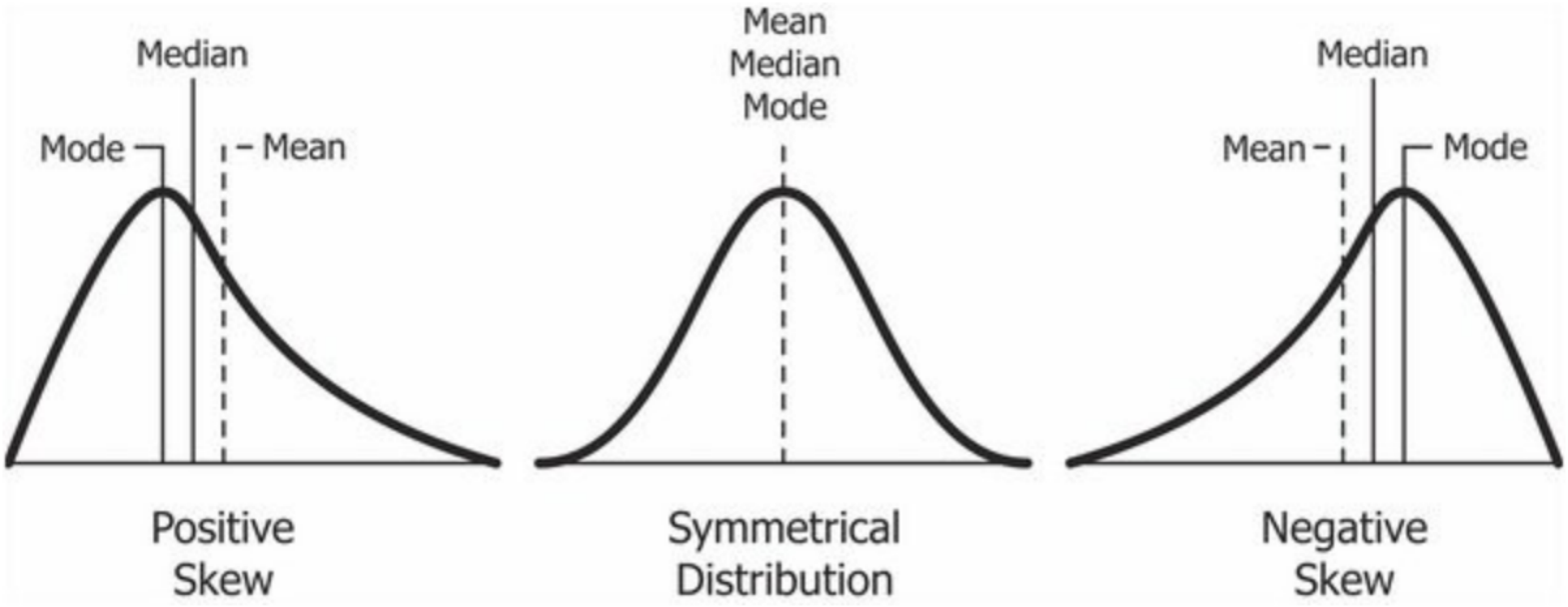
$$\frac{\text{value}_1 + \text{value}_2 + \ldots + \text{value}_n}{\text{sample size}};$$

- **median**: halfway point of the data on the numberline;
- **mode**: location of a peak;

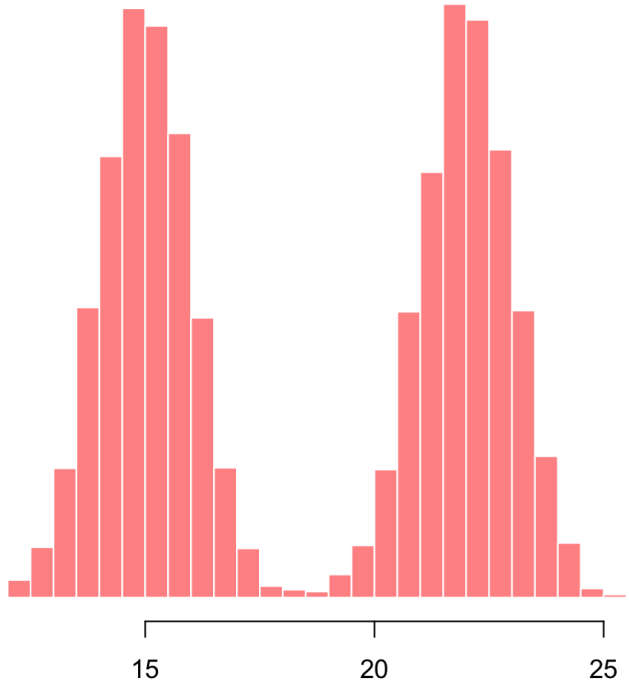
Measures of "spread":

- **standard deviation**

Mean, median, and mode

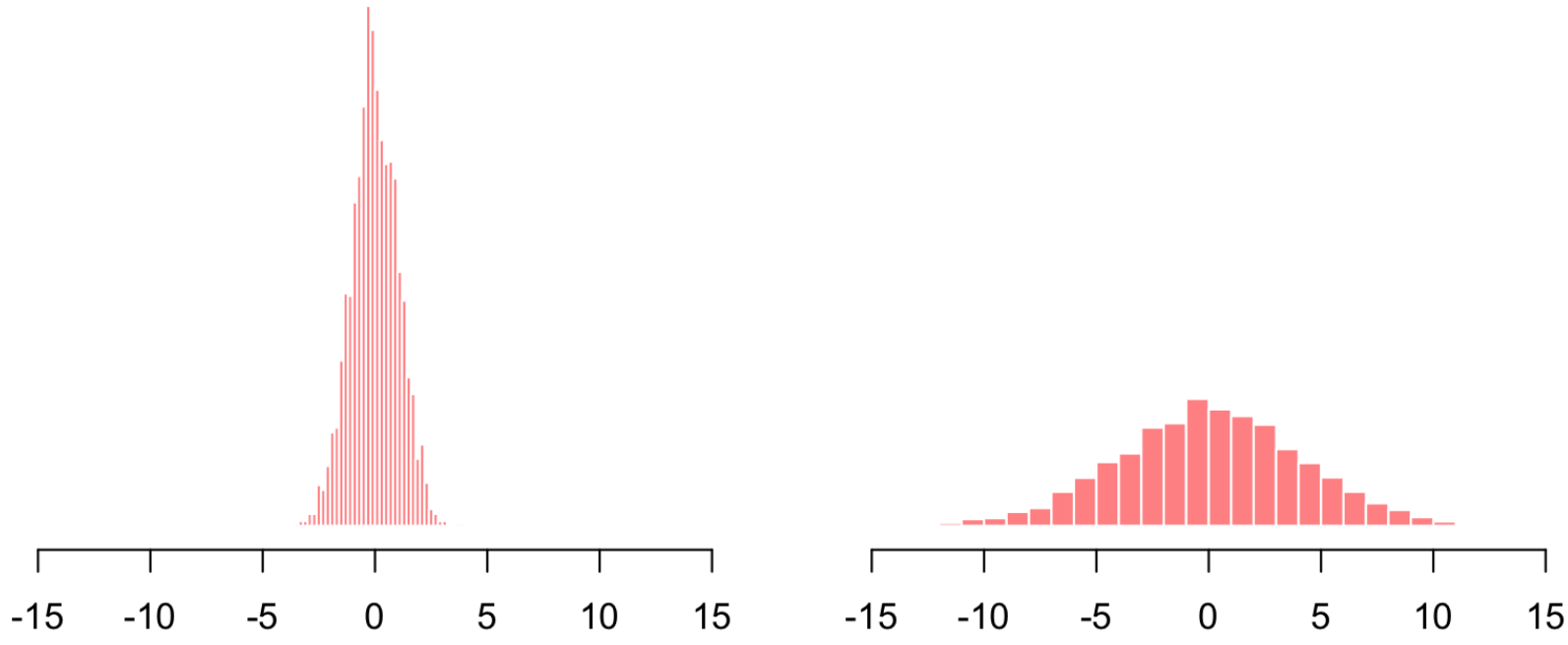


What would be the best guess here?



(This distribution is symmetric, so the mean and median are the same, and they are smack in the middle. But there are not a lot of data values that actually live in the middle, so this is perhaps a poor guess. The two modes are indistinguishable, so at this point we would need some **domain knowledge** to understand why things look this way and come to a conclusion.)

Which dataset provides stronger conclusions?



The one on the right exhibits higher variability (e.g. standard deviation).

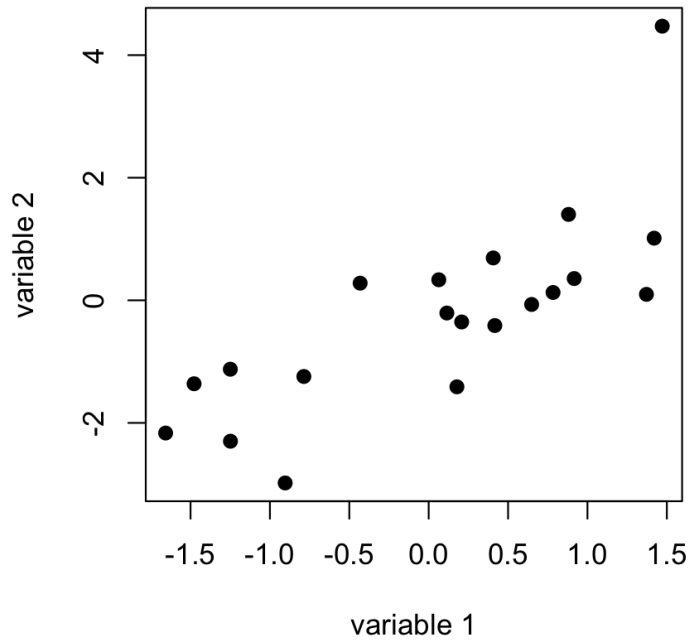
Baby's second data visualization: the scatterplot

Scatterplot: displays the relationship between two variables. For each observation, plot the two variables as an ordered pair in the *xy*-plane.

Turn this...

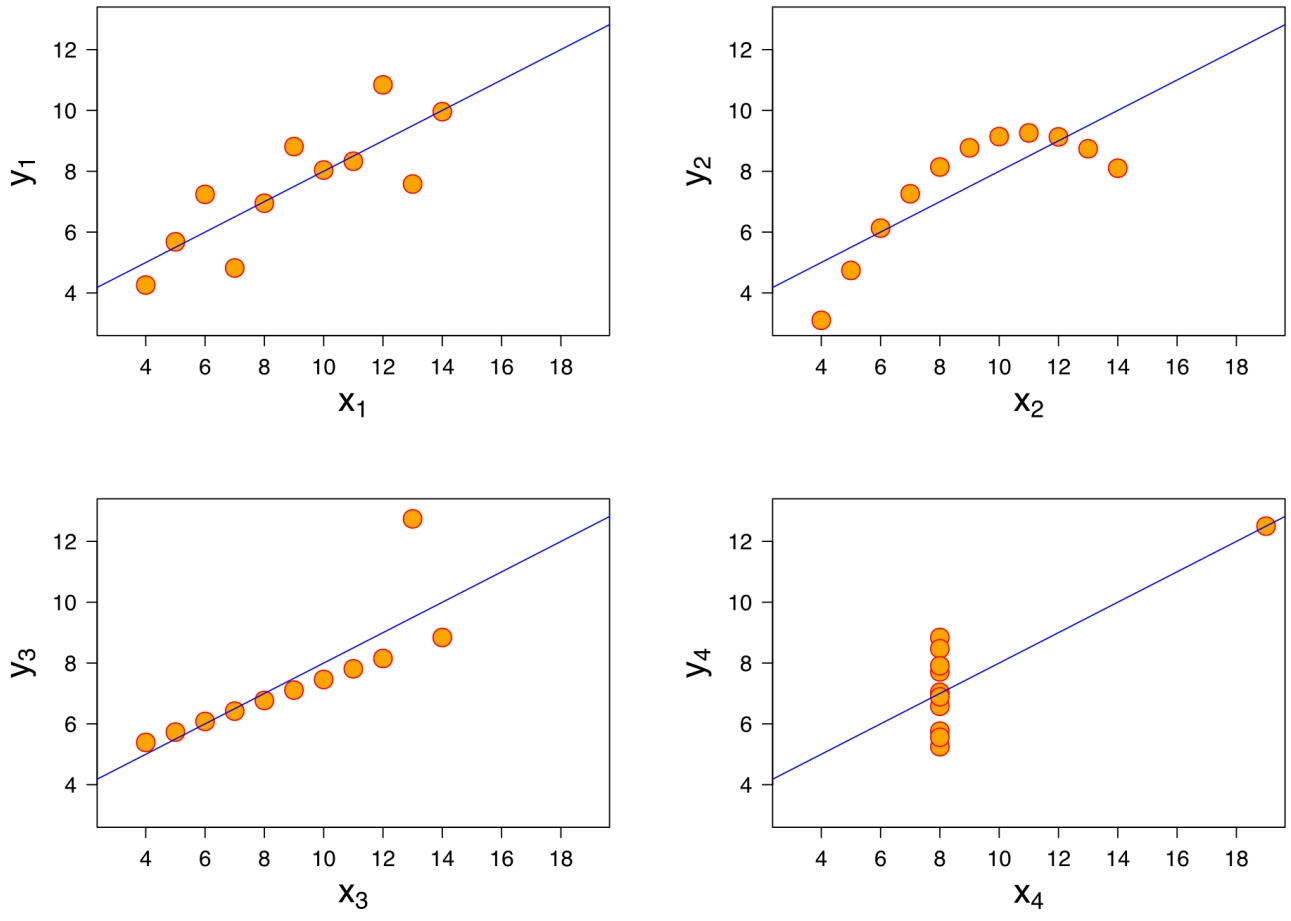
subject	variable_1	variable_2
1	-1.65692830	-2.16524631
2	-0.90396488	-2.97993045
3	1.37141732	0.09720280
4	-0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	-0.78625051	-1.24276055
8	0.64835135	-0.06749005
9	0.06363568	0.33517580

...into this



Anscombe's quartet (the classic version of Lab 1)

These all have the same basic summary statistics:



ABV: **A**lways **B**e **V**isualizing

Causal effect studies

Questions

- Does a medical intervention like a drug or vaccine have an effect? Positive or negative? Big or small?
- Does a policy intervention have an effect, like changes in minimum wage law putting people out of work (or not)?
- Does a change in web design increase site traffic, revenue, etc?

Types of studies

- **Controlled experiment**: researcher randomly assigns treatment vs. control;
    - Example: double-blind clinical trial
- **Observational study**: treatment vs. control were assigned in a messy, real-world way that is related to other factors. Unless you control for those somehow, it is hard to draw causal conclusions.
    - Example: most empirical research in the social sciences (macroeconomics...we cannot run controlled experiments on entire countries)
- **"Natural" experiment**: you stumble upon a setting where nature seems to have randomized treatment vs. control for you. Proceed *as if* a controlled experiment was done (but be very careful!)
    - Example: some empirical research in the social sciences

Natural experiment example

Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania

By DAVID CARD AND ALAN B. KRUEGER*

On April 1, 1992, New Jersey's minimum wage rose from $4.25 to $5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above $5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment. (*JEL* J30, J23)

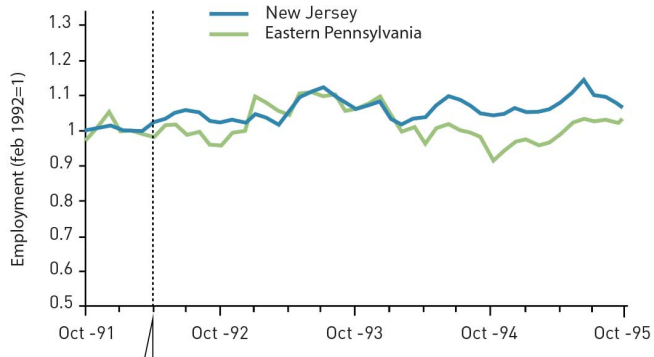
(Communities on either side of the PA/NJ border should be pretty similar. If NJ changes their law but PA does not, which side of the border you happen to fall on might act *like* random assignment. So maybe we can study the causal effect of minimum wage on employment *as if* we ran an experiment in a lab.)

Card and Krueger minimum wage study

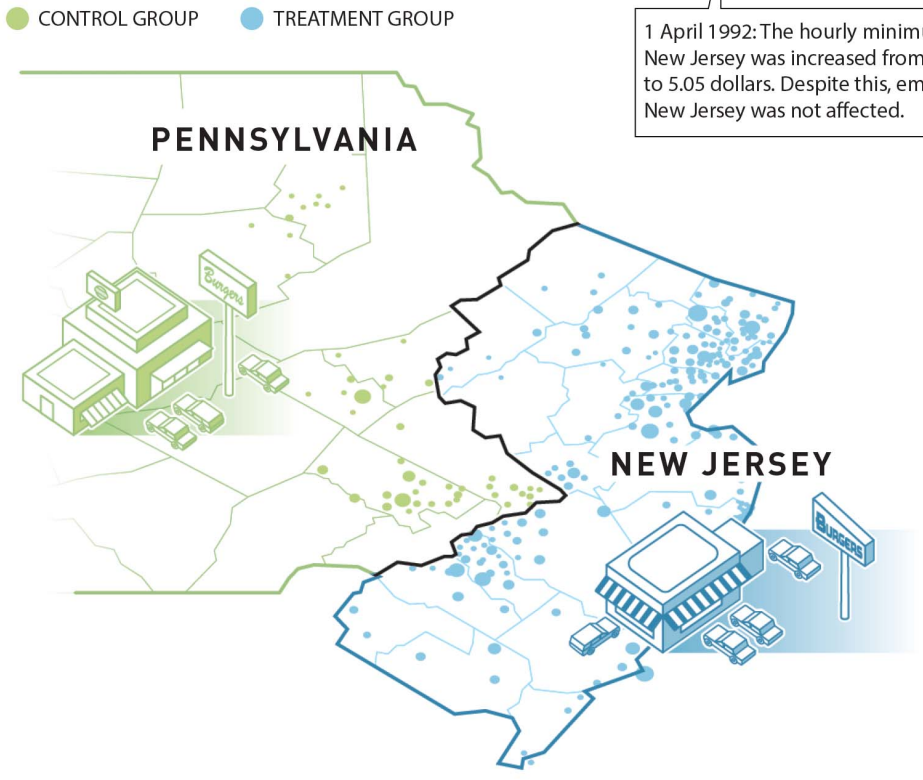
The effect of increasing the minimum wage

Card and Krueger used a natural experiment to study how increasing the minimum wage affects employment.

The researchers identified a treatment group (restaurants in New Jersey) and a control group (restaurants in eastern Pennsylvania) to measure the effect of increasing the minimum wage.



1 April 1992: The hourly minimum wage in New Jersey was increased from 4.25 dollars to 5.05 dollars. Despite this, employment in New Jersey was not affected.

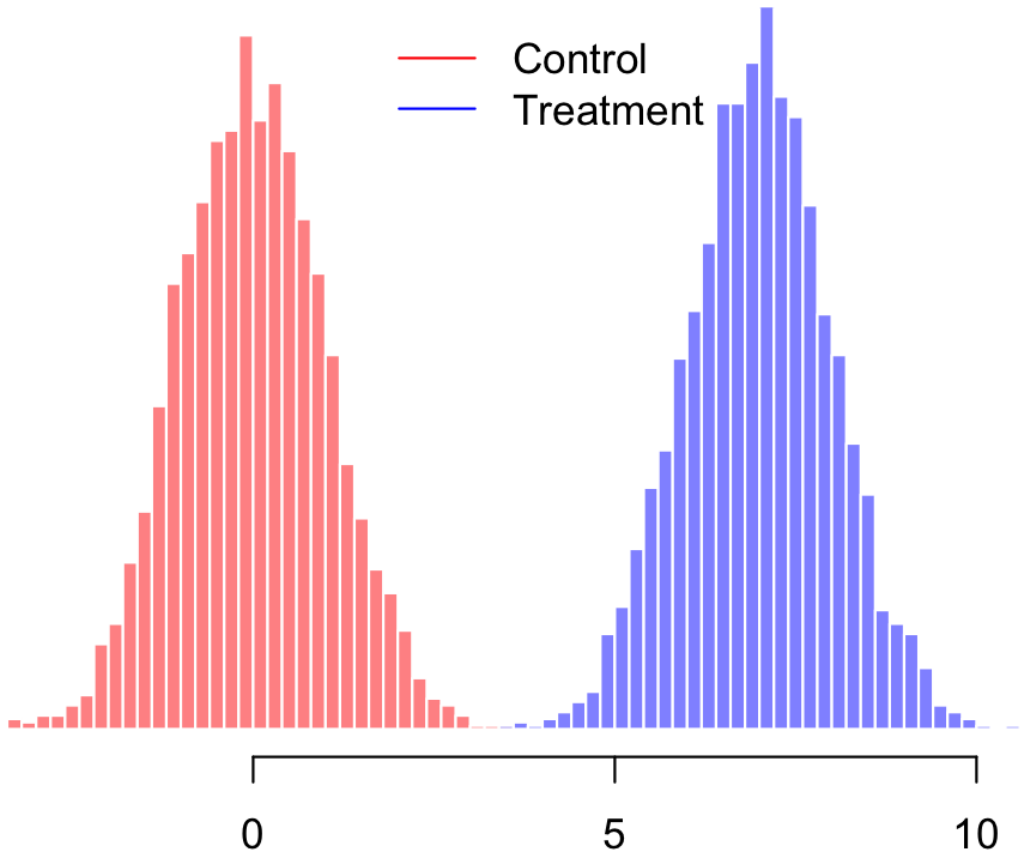


Experimental data

subject	group	response	var1	var2
1	treatment	-0.5719383	-0.05251981	1.16277727
2	treatment	0.5726398	-1.06082751	-1.60894178
3	control	-0.3928624	0.29495654	1.50985796
4	treatment	0.3980274	-0.11390159	1.02469160
5	treatment	0.7439978	0.75729102	-0.83951036
6	treatment	2.3917410	0.13932808	0.03036893
7	control	0.1567981	2.63592181	-0.80432734
8	treatment	1.5417372	0.18982365	0.86441107
9	treatment	-0.9679939	-0.89860125	-0.51709453

Did the treatment have an effect?

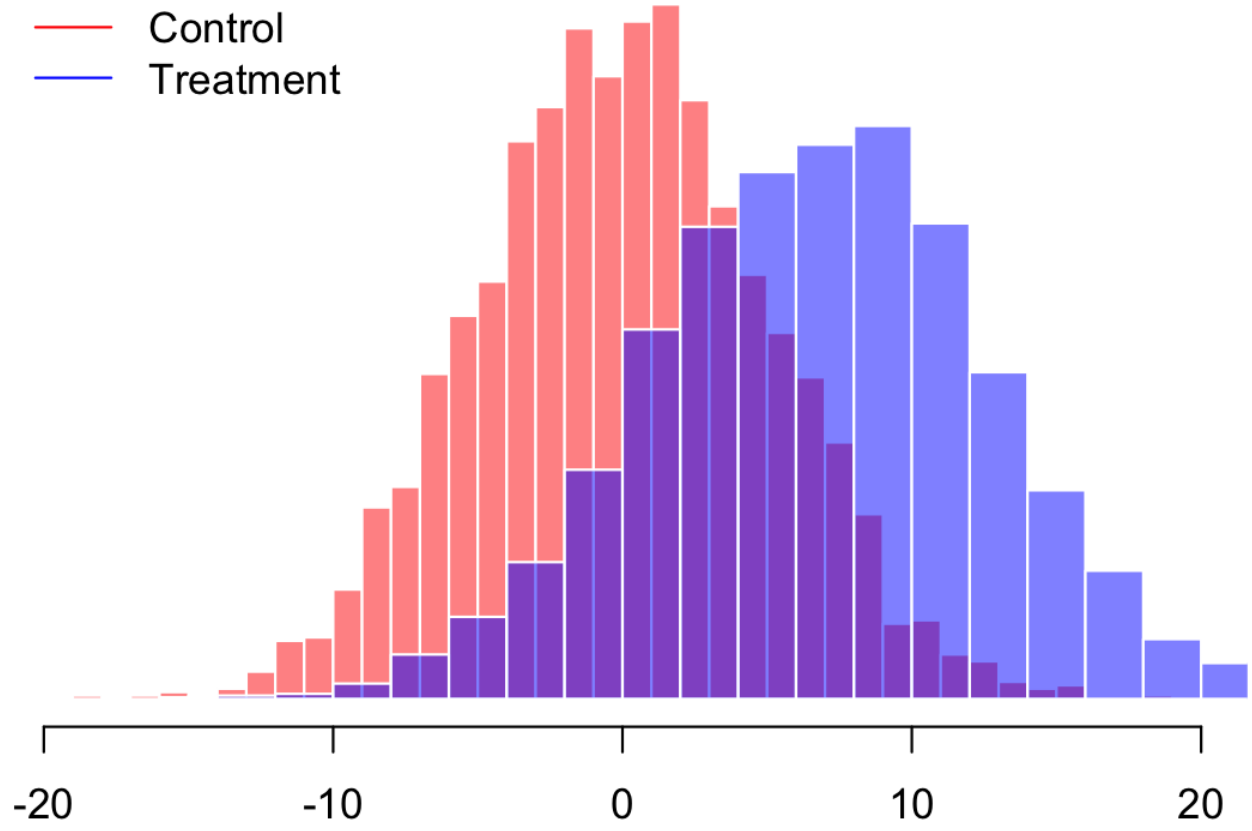
(Assume a controlled experiment was done)



(Probably. These groups are clearly separated. Is the effect big and important? Again, we'd need domain knowledge to know. In this case that would be something basic like **what are the units?**)

Did the treatment have an effect?

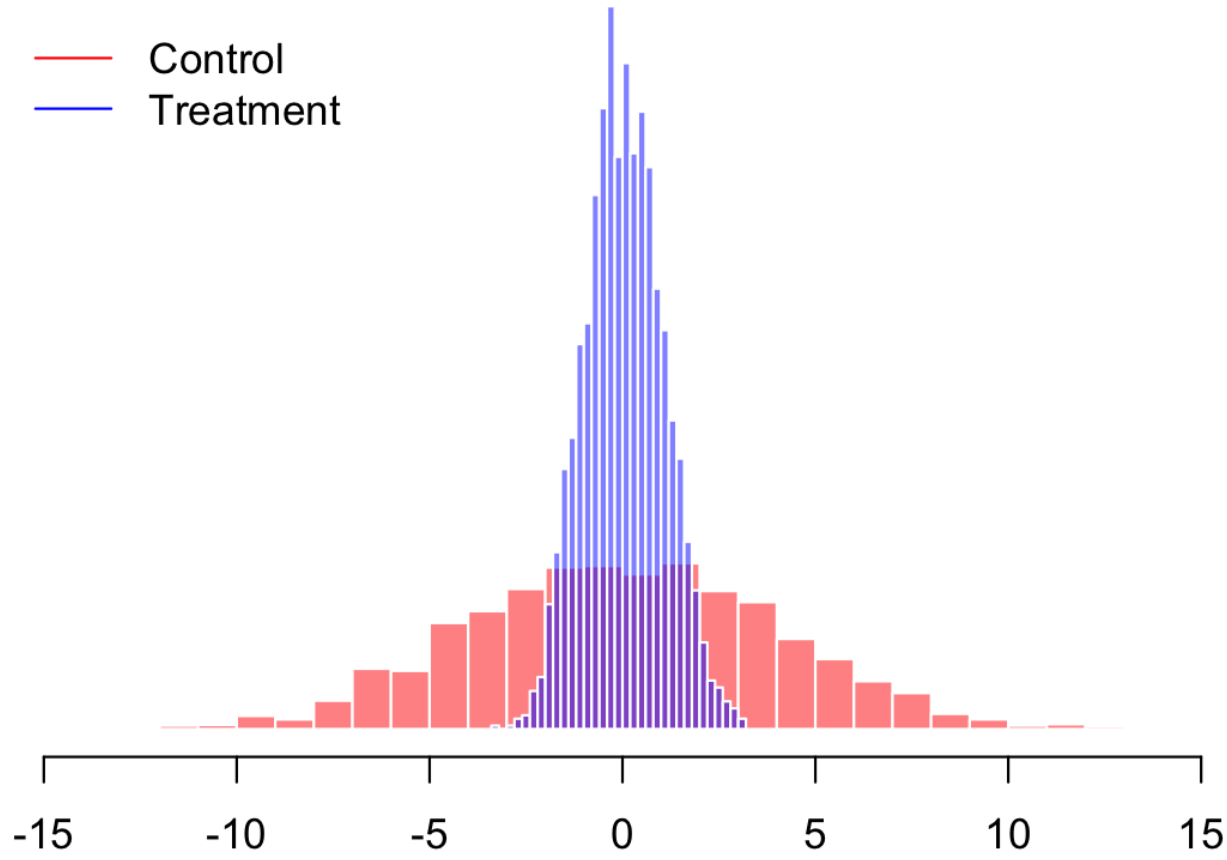
(Assume a controlled experiment was done)



(The best guess is that it probably did, but there's clearly more uncertainty here than in the previous case. Increased variability in the data corresponds to less certain answer to the question you asked.)

Did the treatment have an effect?

(Assume a controlled experiment was done)



(Probably yes, but in a different way than the last two examples. The effect of the treatment is not to shift the location of the distribution, but rather to tighten the variance.)