Section 2: summarizing single variables

STOR 155.02, Spring '21

updated 2021-01-30

What you will learn

- sample distributions
- plots: bar graphs, histograms
- centrality statistics: mean, median
- spread statistics: range, percentiles (aka quantile), variance and standard deviation

Resources

- Textbook ch 1.6.2-1.6.5 and 1.10

Looking at data

Definition: sample distribution

> The sample distribution of a variable tells us what values it takes and how often it takes these values in our dataset.

this is about the values a variable actually has in your data

does *not* tell you about all possible values your variable could have if you got more or different data

frequency statistics

- **frequency** of a category is the **number of observations** in that category
- **relative frequency** is number of **observations in a category / total** number of observations

distributions: categorical

- given by **relative frequencies** of each category
- think of it as the chance a randomly chosen observation falls in that category
- e.g. relative frequency of 0.1 for category 'A' means 10 percent of observations of type 'A'

distributions: quantitative

- given by **relative frequencies in a range** of possible values
- think of as the chance a randomly chosen observation falls in the range
- will evaluate with **histogram** plots

Looking at sample distributions: depends on the variable type

Categorical:

bar graphs

pie charts

Numeric/quantitative:

histograms

Looking at categorical data

bar graph

- bar height represents frequency or relative frequency for each category
- the frequency must first be computed

pie chart

- area of a pie slice represents the relative frequency of a category
- these can be **hard to read** because areas are difficult to estimate by eye
- can be used **only for** counts adding to the total number of observations, or percentages adding to 100

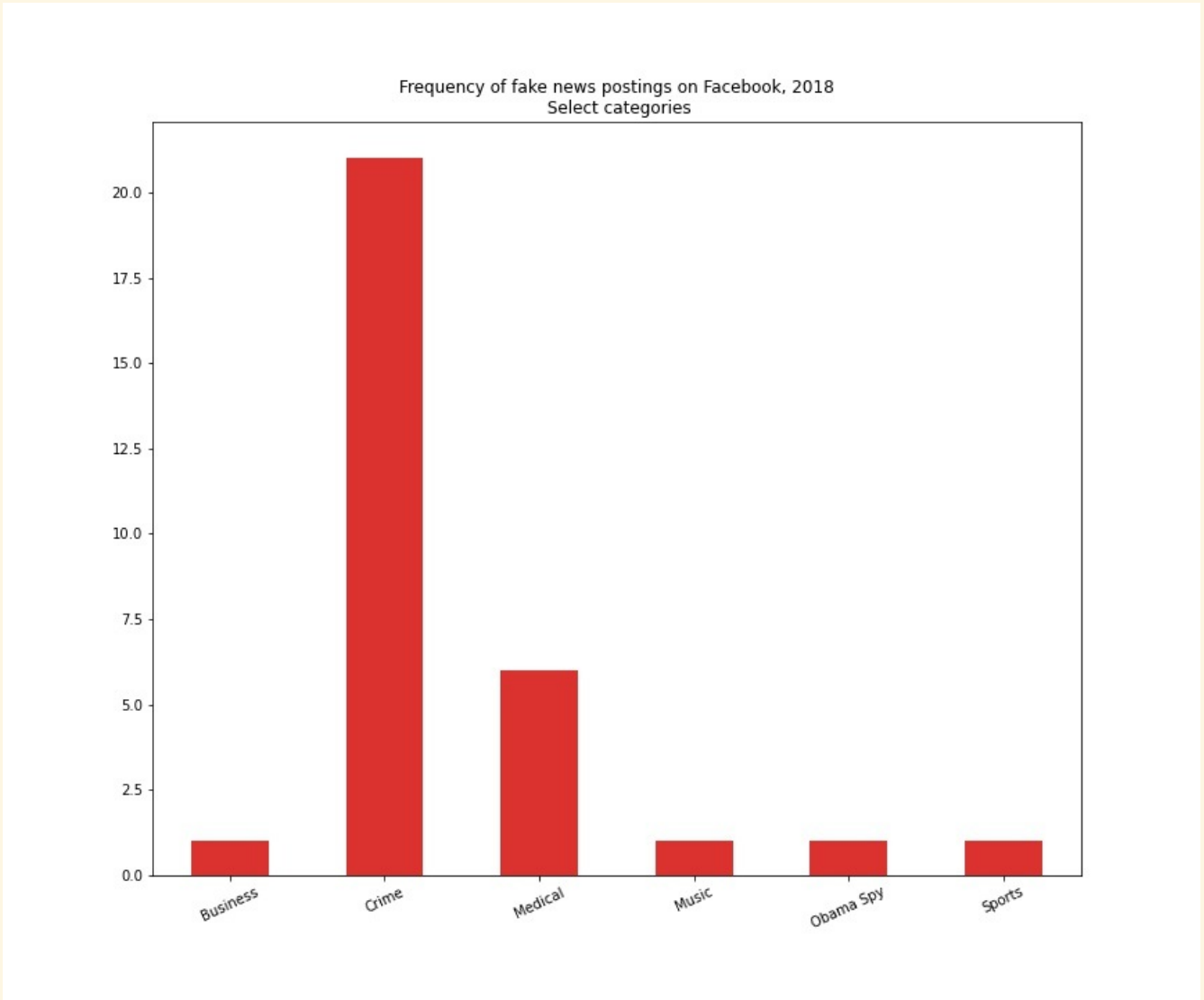
Example: Fake news on facebook

dataset

- data on Facebook popularity of posts by fake news websites in 2018
- identified and collected by Buzzfeed News
- Biggest fake news hits on Facebook in 2018
- source: Buzzfeed github site
- dataset on our course github page as well

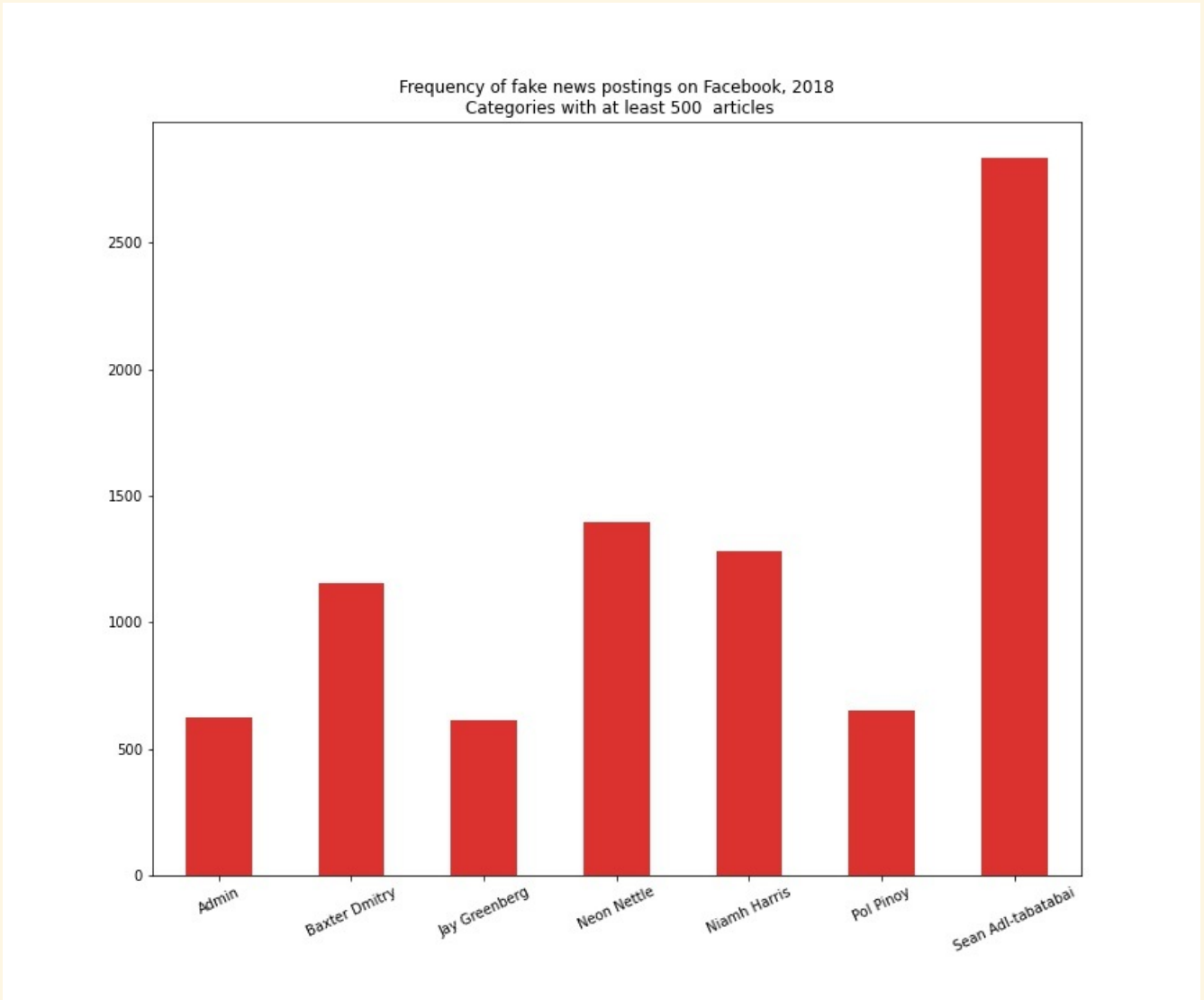
observations and variables

- each row is represents an article
- `fb_engagement` is essentially number of likes post received
- `category` is a buzzfeed-created categorical variable


Frequency of fake news postings on Facebook, 2018
Select categories
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Business
Crime
Medical
Music
Obama Spy
Sports

There are too many weird small categories to show nicely, so I picked some I was interested in.

but what does it look like if we isolate the most common posting categories?


Frequency of fake news postings on Facebook, 2018
Categories with at least 500 articles
2500
2000
1500
1000
500
0
Admin
Baxter Dmitry
Jay Greenberg
Neon Nettle
Niamh Harris
Pol Pinoy
Sean Adl-tabatabai

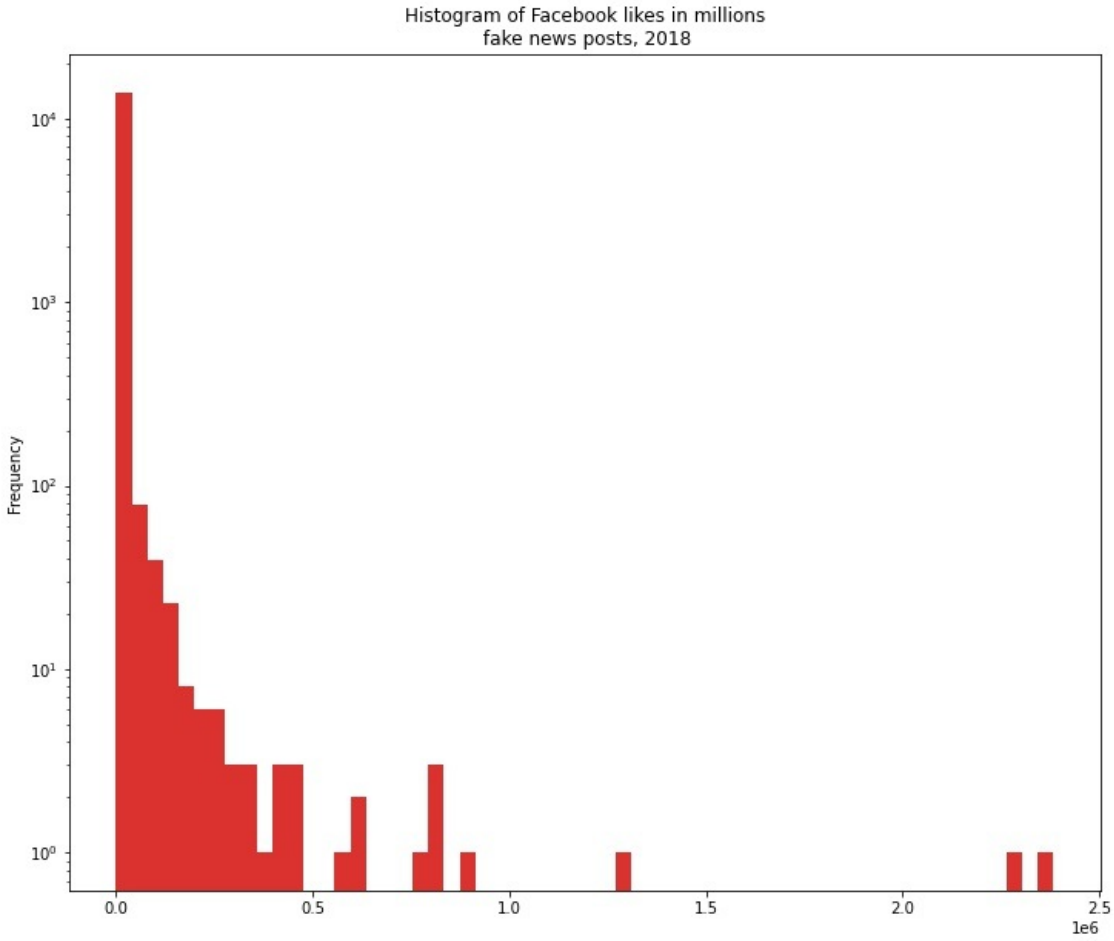
Does this information make sense?

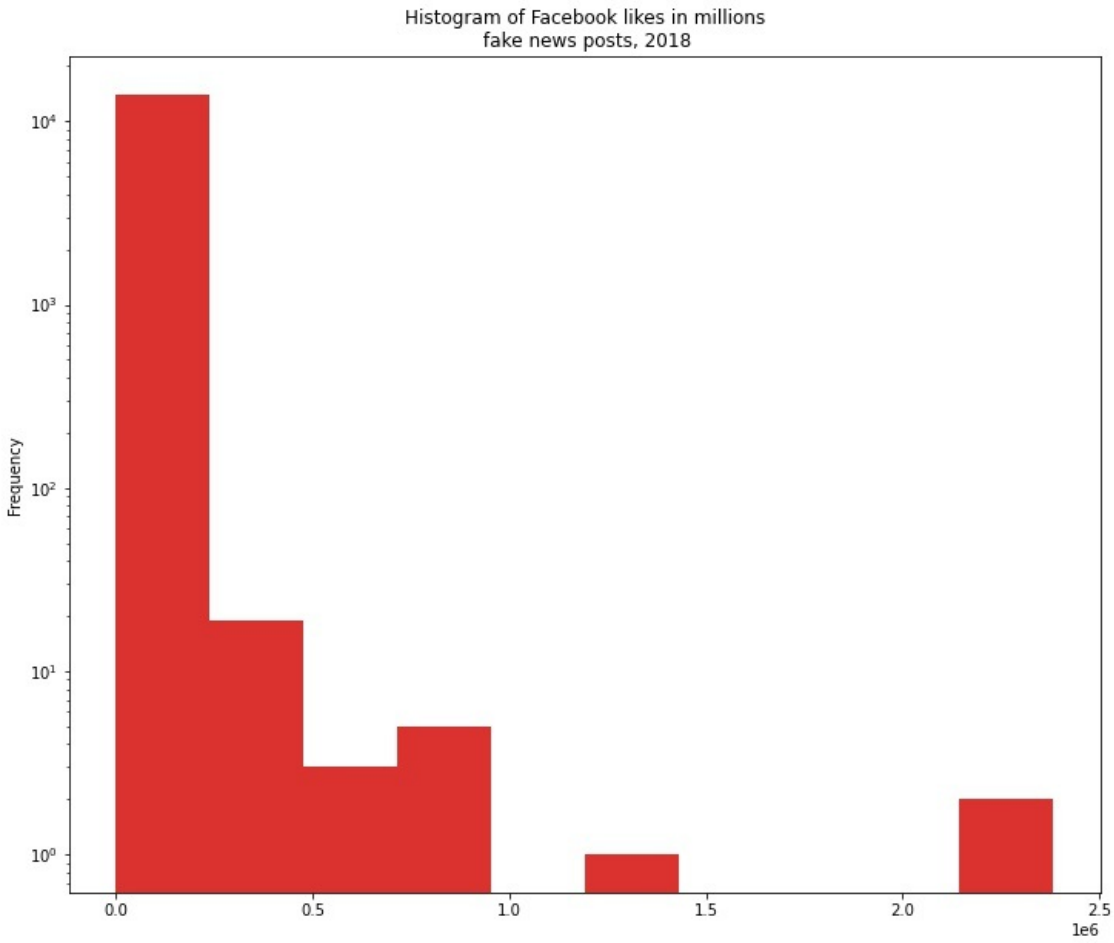
Will investigate a little in the group exercise.

Histogram: looking at quantitative variable distributions

- **breaks the range of the values into bins** (intervals) of a given width
- displays frequencies or relative frequencies of the observations that fall into each interval
- choose any convenient number of intervals of equal width
- but choice of width might affect how you interpret the data

Example: `fb_engagement` variable histogram with differing bin sizes


Histogram of Facebook likes in millions
fake news posts, 2018
Frequency
10^4
10^3
10^2
10^1
10^0
0.0
0.5
1.0
1.5
2.0
2.5
1e6


Histogram of Facebook likes in millions
fake news posts, 2018
Frequency
10^4
10^3
10^2
10^1
10^0
0.0
0.5
1.0
1.5
2.0
2.5
1e6

A good histogram allows you to guess at key properties of a numeric variable:

1. What are the maximum and minimum values?
2. What is the 'middle' value?
3. What is the average value?
4. Are the values concentrated around the average or spread out?
5. What is (are) the most common value(s)?

These are all important *statistics* of a numeric variable.

Let's make these ideas more precise

Sample statistics for numeric variables

Range:

- max value minus min

Nth percentile (aka quantile):

- value defining a cutoff so that N percent (appropximately) of your data is below that value

Median:

- 50th percentile, half of the data is below this value, half above

Mean:

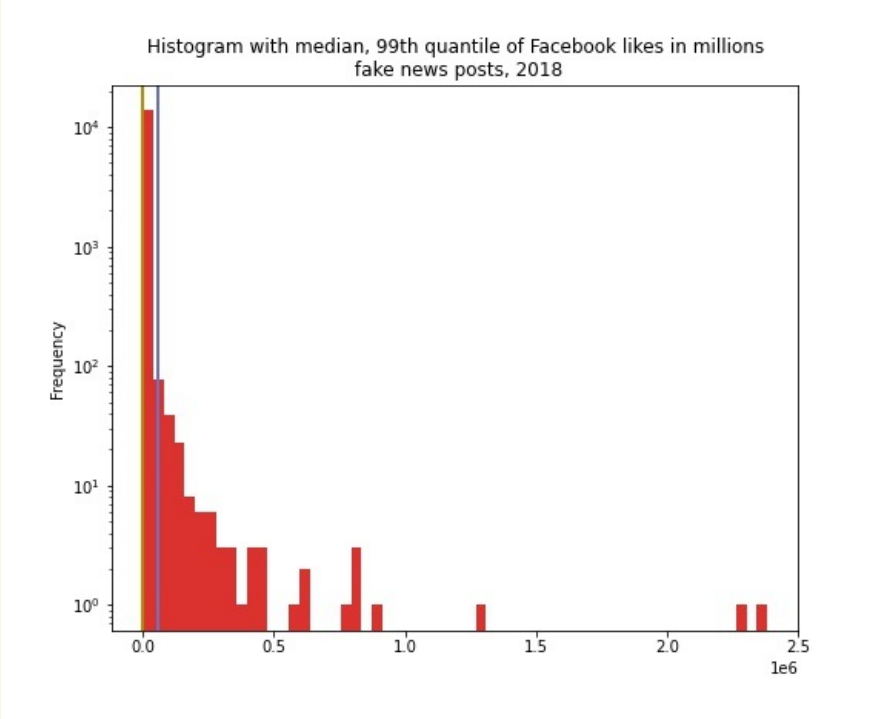
- average

Variance:

- average squared distance of values from the mean. A way to think about how *spread out* the data are

Example: Guessing the Nth percentile

Use bar height to guess bin such that N percent of data is below it. Hard to tell with some plots! Remember $1e6 = 10^6 = 1$ million.



Math definitions

If I have some number n data points, I will represent my numeric variable values by

$$X_1, X_2 \ldots X_n$$

mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots X_n$$

variance

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

Warning!

For reasons not important to this class, sometimes statistical software will calculate variance with $\frac{1}{n-1}$ in front instead of $\frac{1}{n}$.

We'll talk about it in the data homework.

Averages are everywhere. What do they tell us, and not?

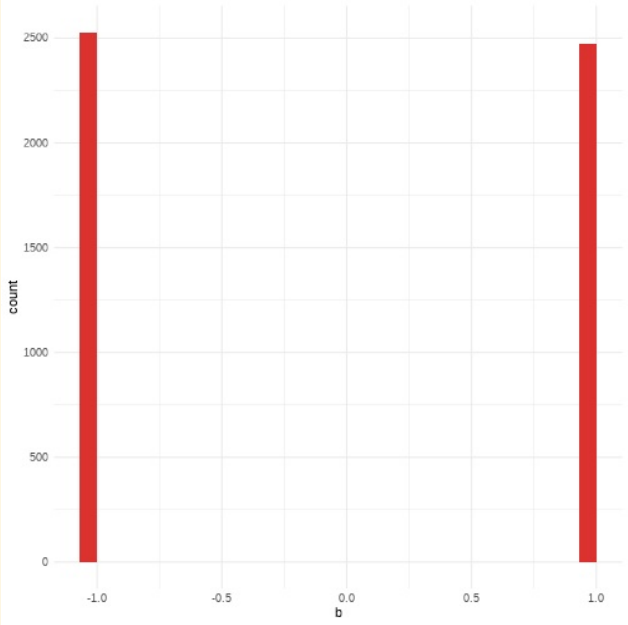
does not tell you

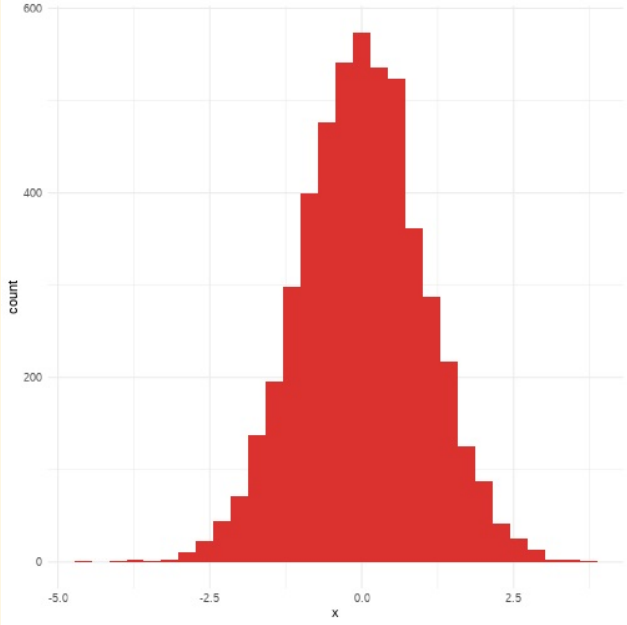
- how evenly values are spread
- min/max
- whether a few extremely big (positive or negative) values are influencing the calculation
- **median** tells you about the middle but **is less sensitive to extreme values**

but important because

- a common way to capture the 'middle' value, along with median
- magic of the bell curve: we'll see this later!

These two variables have approximately the same average, $\bar{X} = 0$





Group breakout exercise

Poll closes at __!

Poll everywhere link:
https://PollEv.com/surveys/wk1afrjltBXR1J1GQIO

You must be registered to submit a response!

Discuss with group but respond individually


Five more minutes