CSCI 3022

intro to data science with probability & statistics

August 29, 2018

1. Exploratory data analysis
2. Summary statistics

- ☐ github
- ☐ moodle
- ☐ piazza ←
- ☐ python ?

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Populations and Samples

Data scientists hope to learn about some characteristic/variable of a population

But we can't actually see or study the whole population, so we investigate a sample.

- **Definition**: A population is a collection of units (people, songs, shoes, pandas).
- **Definition**: A sample is a subset of the population.
- **Definition**: A characteristic/variable of interest (VoI) is something we want to measure for each unit.

Populations and Samples

Data scientists hope to learn about some characteristic/variable of a population

But we can’t actually see or study the whole population, so we investigate a sample

Populations and Samples

Data scientists hope to learn about some characteristic/variable of a population

But we can't actually see or study the whole population, so we investigate a sample

Example: Suppose the city of Denver wants to estimate its per-household income via a phone survey. They call every 50th number on a list of Denver phone numbers between 6pm and 8pm. In this case, what is

Populations and Samples

Data scientists hope to learn about some characteristic/variable of a population

But we can't actually see or study the whole population, so we investigate a sample

Example: Suppose the city of Denver wants to estimate its per-household income via a phone survey. They call every 50th number on a list of Denver phone numbers between 6pm and 8pm. In this case, what is

- the population? Households of Denver
- the sample? Every 50^{th} ~~person~~ #
- the variable of interest? Household Income

Populations and Samples

Data scientists hope to learn about some characteristic/variable of a population

But we can't actually see or study the whole population, so we investigate a sample

Example: Suppose the city of Denver wants to estimate its per-household income via a phone survey. They call every 50th number on a list of Denver phone numbers between 6pm and 8pm. In this case, what is

- the population?
- the sample?
- the variable of interest?

Phone Book

Definition: The sample frame is the source material or device from which sample is drawn.

Sample Types

- **Simple (uniform) random sample**: randomly select people from the sample frame. No preference given to anyone in particular.
- **Systematic sample**: order the sample frame. Choose integer *k*. Sample every *k*th unit in the sample frame.
- **Census sample**: sample literally everyone in the population.
- **Stratified sample**: suppose you have a *heterogenous* population that can be broken up into *homogenous* groups. Randomly (uniformly) sample from each group proportionate to its prevalence in the population.

60%

40%

think: bike to class vs walk to class

Example: What type of sample was done in the previous example of per-household income phone calls?

Populations and Samples

Data scientists hope to learn about some characteristic or variable of a population by studying a sample.

A major part of this course is about how you can make the jump from studying a sample to drawing conclusions about the population. And not just how, but when… and why!

This process is called *inference*.

Exploratory Data Analysis EDA

Before we learn about inference, we're first going to learn how to explore the data.

This is useful for summarizing and recognizing patterns in the data, or comparing one dataset to another.

There are two main types of of data exploration: **Numerical** and **Graphical**

Exploratory Data Analysis

Before we learn about inference, we're first going to learn how to explore the data.

This is useful for summarizing and recognizing patterns in the data, or comparing one dataset to another.

There are two main types of of data exploration: **Numerical** and **Graphical**

Numerical summaries are exactly what they sound like: ways of summarizing a whole dataset using numbers.

Calculating and interpreting certain numerical summaries of a sample can help us gain a better understanding of what's going on in that sample, so we call these numerical summaries of a sample **sample statistics**.

Measures of Centrality

Summarizing the "center" of the sample data is a popular and important characteristic of a set of numbers.

Goal: Capture something about the "typical" unit in the sample with respect to the VoI.

There are three popular measures of the center of a sample:

mean: average. If choosing from sample uniformly at random then mean is the expected value of VoI.

median: exact middle value. (person in middle)

mode: most popular value. Appears most often.

The Sample Mean

For a given set of numbers $x_1, x_2, x_3, \ldots, x_n$, the most familiar measure of of the center is the mean (also called the *arithmetic average*).

Definition: The sample mean of observations $x_1, x_2, x_3, \ldots, x_n$ is given by

$$\frac{1}{n}\left[x_1 + x_2 + x_3 + \ldots + x_n\right] = \boxed{\frac{1}{n}\sum_{k=1}^{n} x_k}$$

# of samples.

Example: Compute the sample mean of data 2, 4, 3, 5, 6, 4

$$\frac{1}{6}\left[2+4+3+5+6+4\right]$$

$$= \frac{1}{6}\left[6+8+10\right] = \frac{1}{6}\left[24\right] = 4$$

The Sample Mean

For a given set of numbers $x_1, x_2, x_3, \ldots, x_n$, the most familiar measure of of the center is the mean (also called the *arithmetic average*).

Definition: The sample mean of observations $x_1, x_2, x_3, \ldots, x_n$ is given by

Example: Compute the sample mean of data 2, 4, 3, 5, 6, 40

Advantages of mean: Easy. Intuitive
Disadvantages of mean: Outliers or Errors can distort the mean

The Sample Median

For a given set of numbers $x_1, x_2, x_3, \ldots, x_n$, the **sample median** is the "middle" value when we order the numbers from smallest to largest.

Definition: The sample median of *ordered* observations $x_1, x_2, x_3, \ldots, x_n$ is given by the middle item. This means item $(n+1)/2$ if n is odd, and the mean of items $n/2$ and $(n+1)/2$ if n is even. [*Why*?]

→ we want the middle value.

Example: Compute the sample median of data: ~~36~~, ~~15~~, ~~39~~, ~~41~~, ~~40~~, ~~42~~, ~~47~~, ~~49~~, ~~7~~, ~~6~~, ~~43~~.

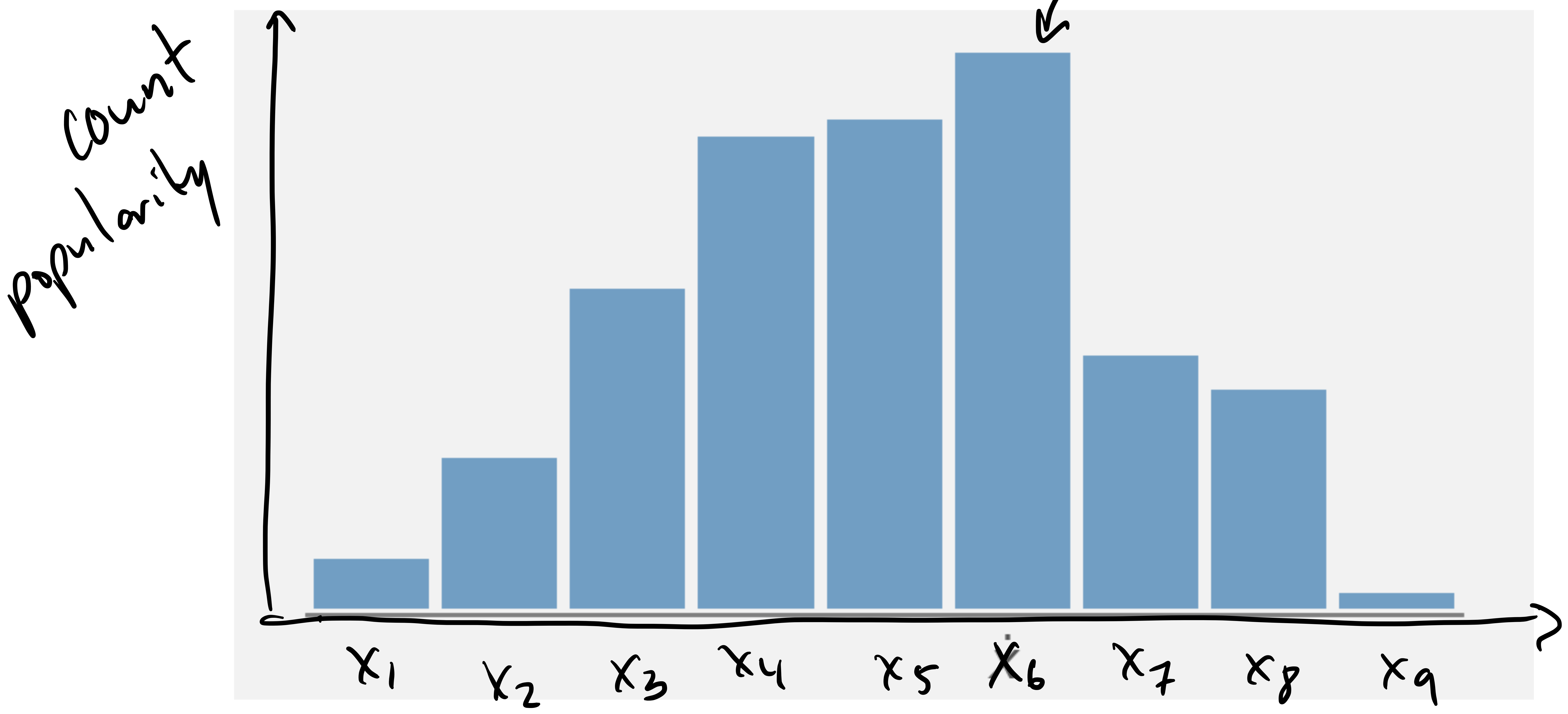
6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

1 2 3 4 5 6 7 8 9 10 11

11 is odd → $\frac{n+1}{2} = \frac{11+1}{2} = 6$

The Sample Mode

Definition: for a given set of numbers $x_1, x_2, x_3, \ldots, x_n$, the **sample mode** is the most popular value.



Challenge:

Come up with three *different* datasets of 5 integers each:

Give me dataset 1 with a mean of 6.

6, 6, 6, 6, 6

26, 1, 1, 1, 1

34, -1, -1, -2, 0

Give me dataset 2 with a median of 5.

5, 5, 5, 5 5

-a billion, -1, 5, 12, 50

Give me dataset 3 with a mode of 8.

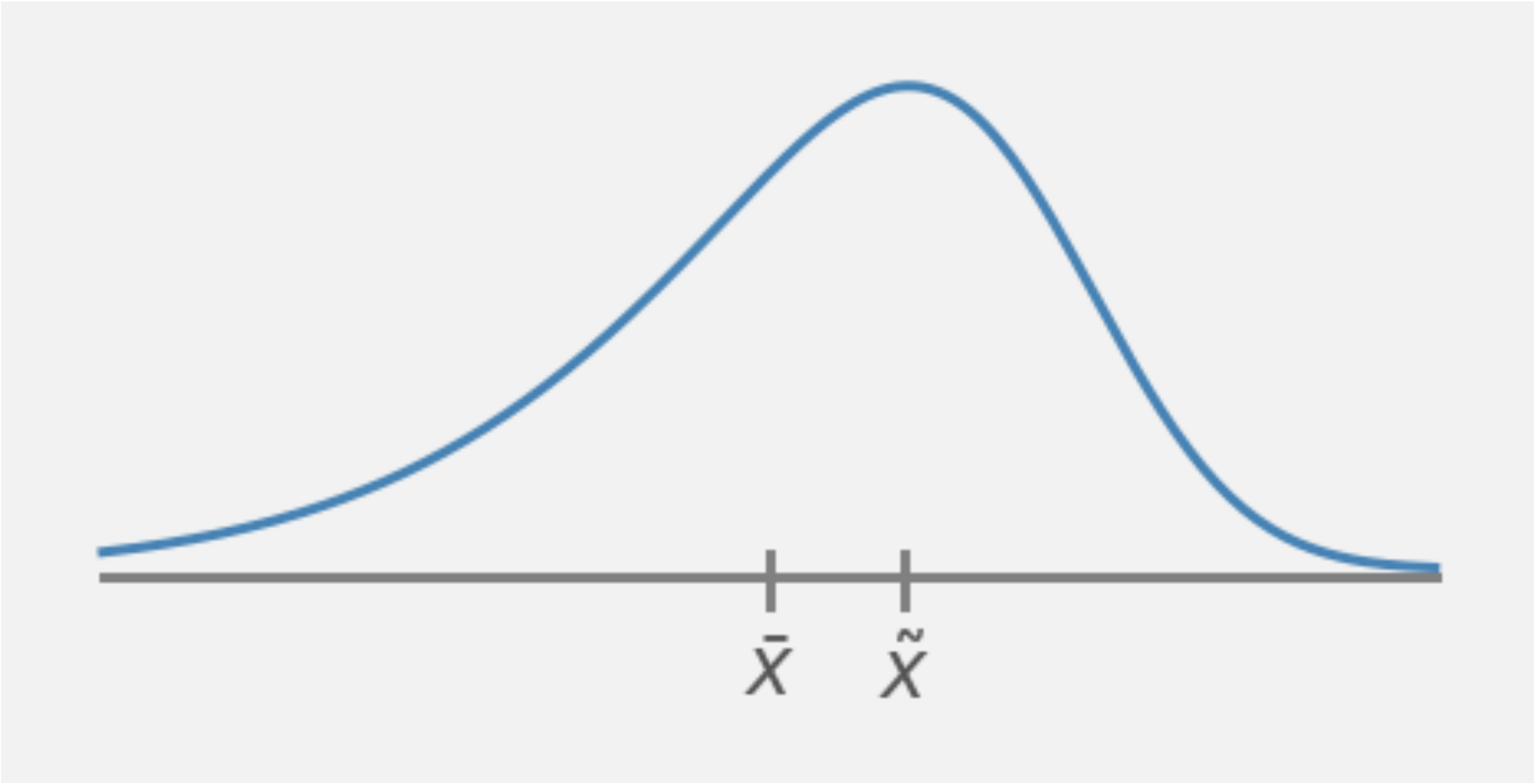
8, 8, 8, 8, 8

8, 0, 8, 9, 12

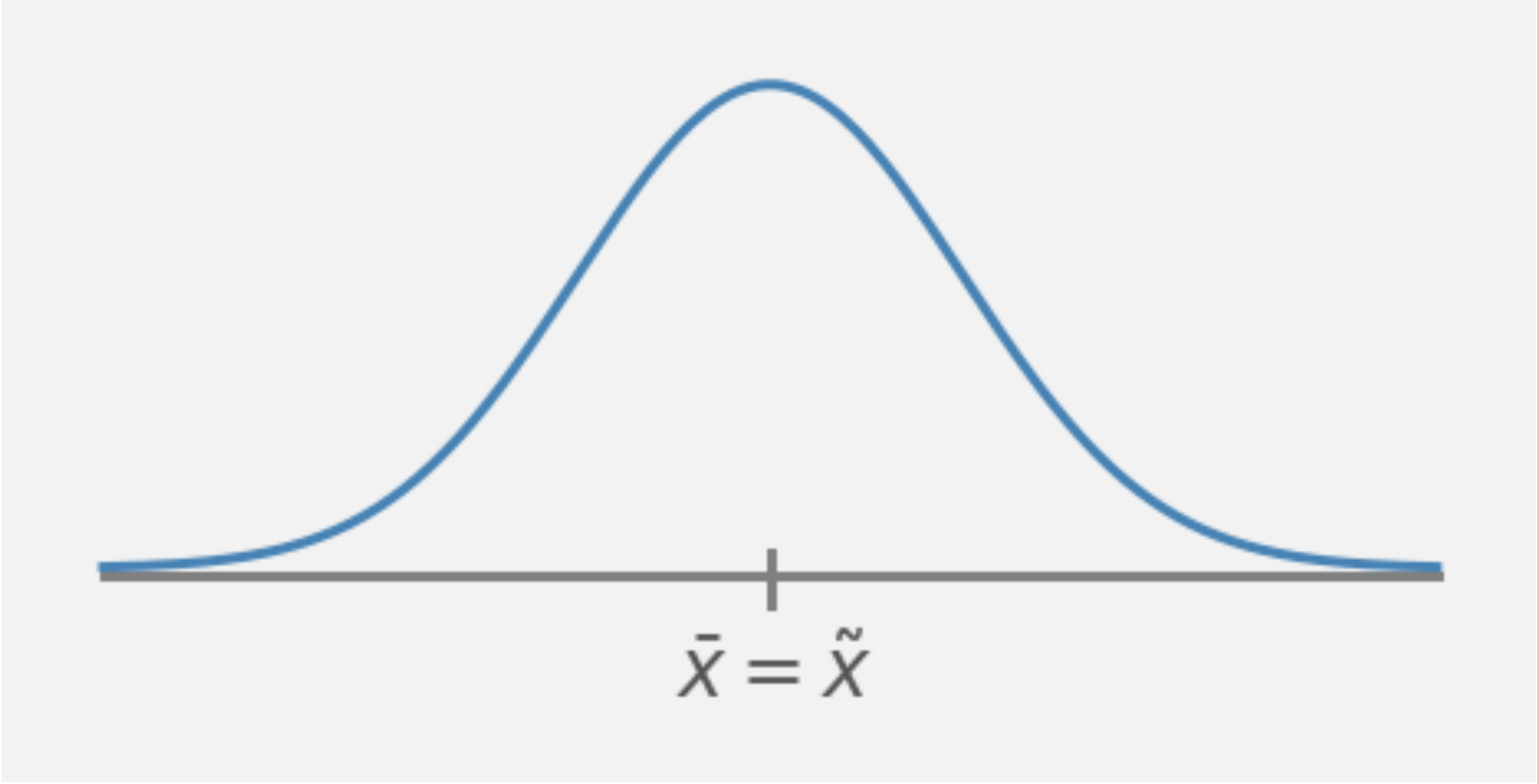
Mean vs. Median

$\bar{x}$ $\tilde{x}$

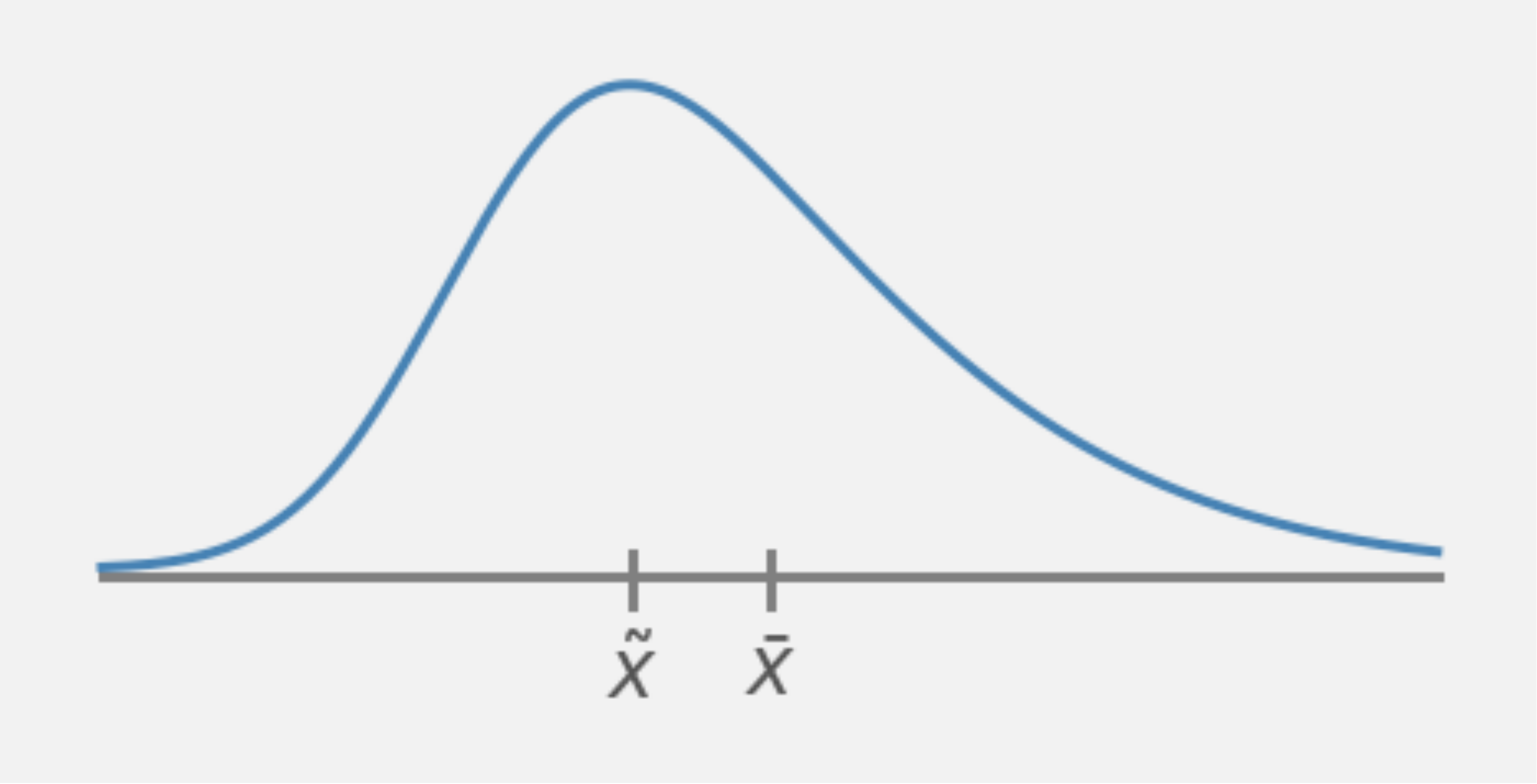
The population mean and median will not, in general, be identical. If the population distribution is positively or negatively **skewed**…



mean < median



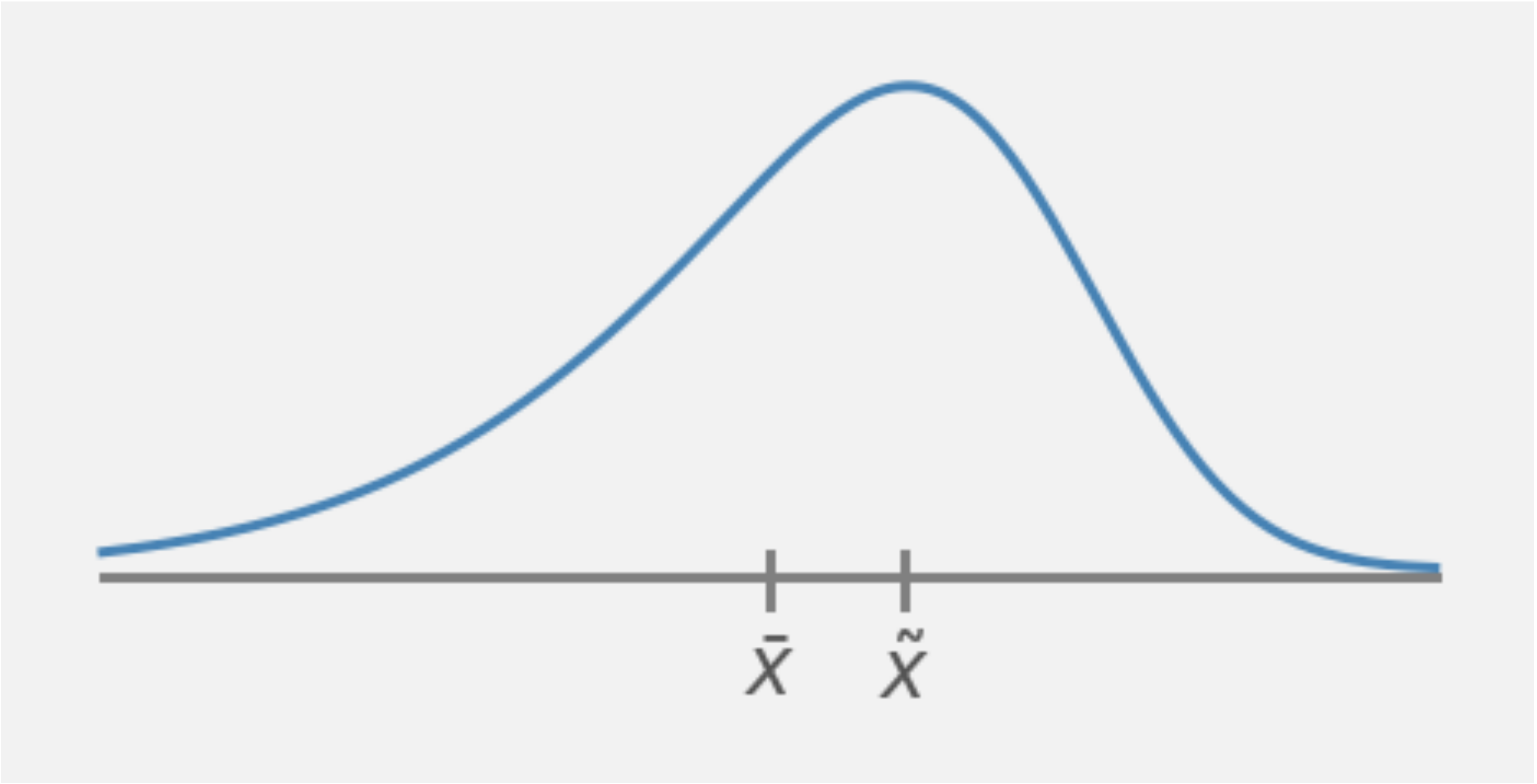
no skew



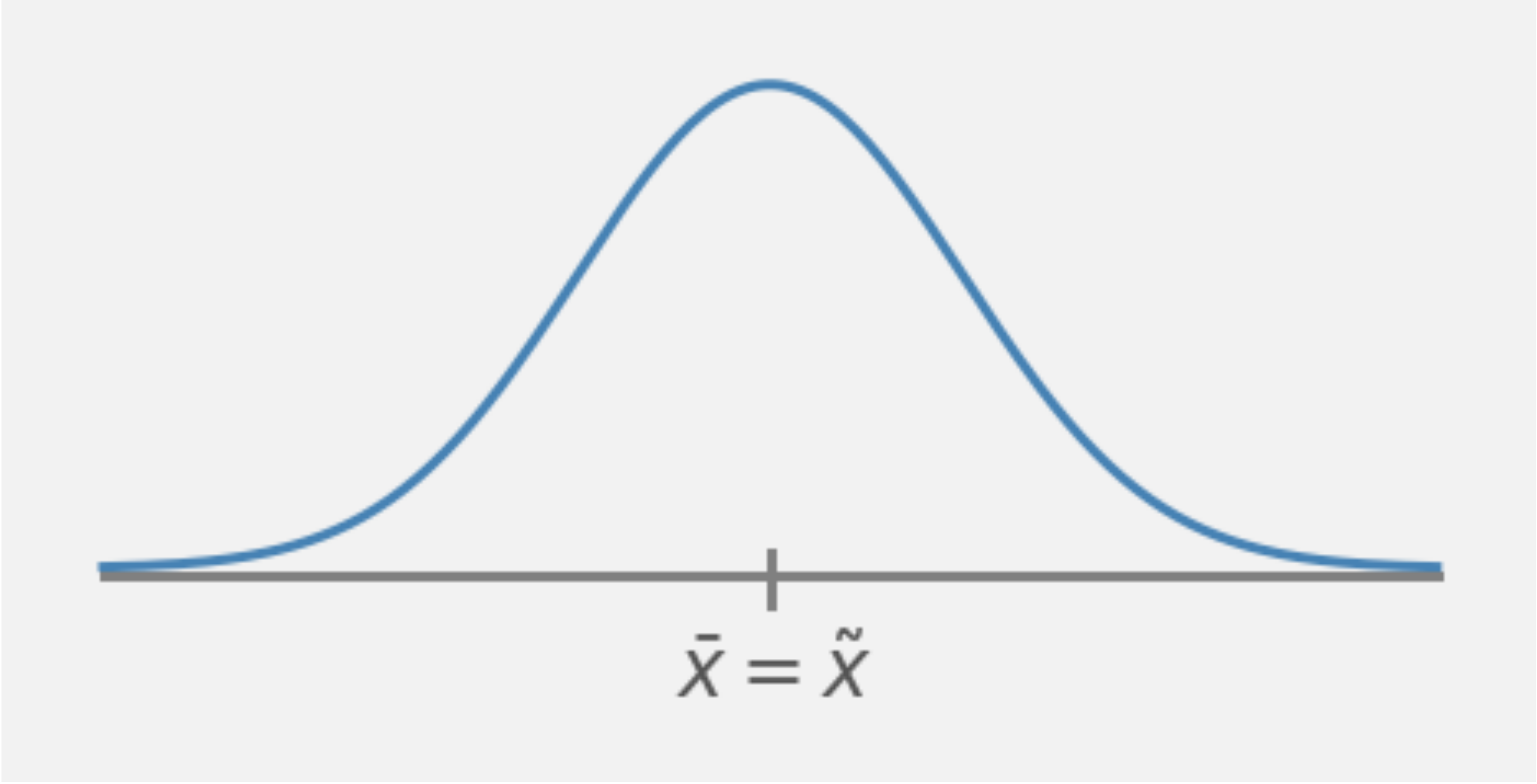
mean > median

Mean vs. Median

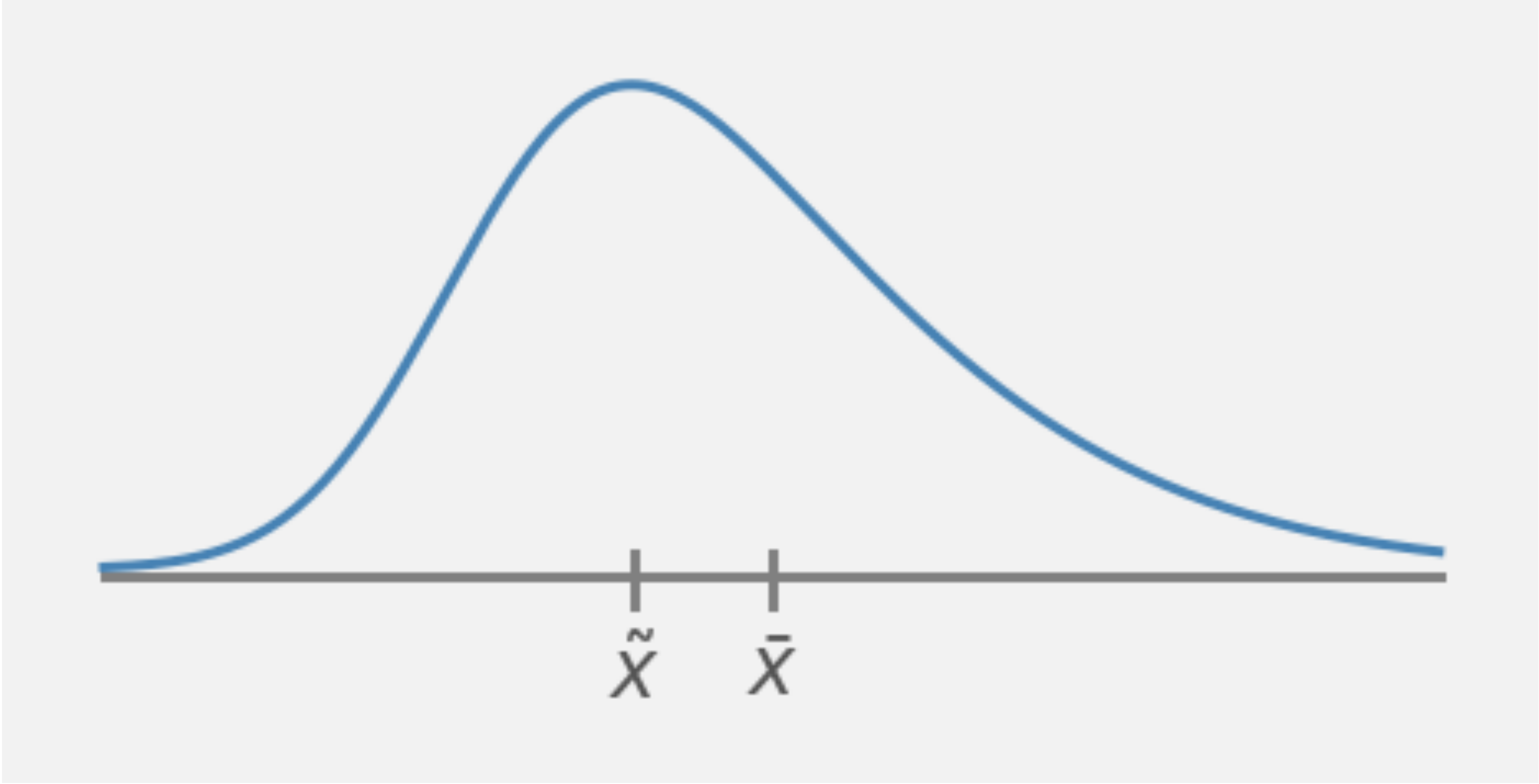
The population mean and median will not, in general, be identical. If the population distribution is positively or negatively **skewed**…



mean < median
negative skew
left skew



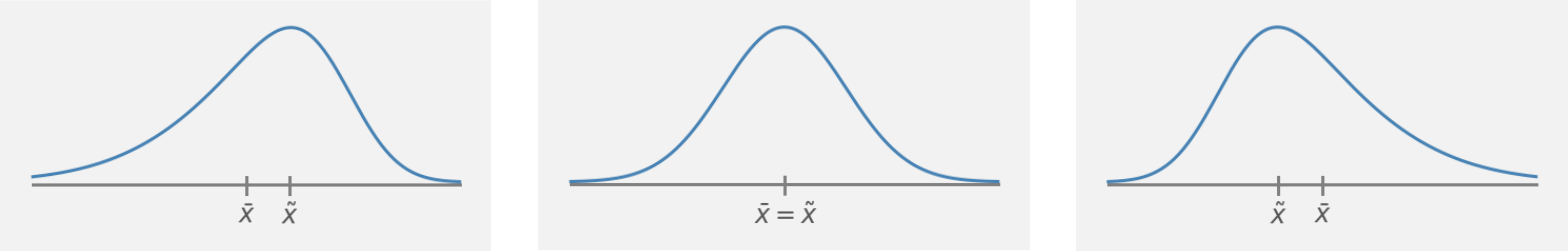
no skew



mean > median
positive skew
right skew

Mean vs. Median

The population mean and median will not, in general, be identical. If the population distribution is positively or negatively **skewed**…



mean < median	no skew	mean > median
negative skew		*positive* skew
left skew		*right* skew

bears, beets, battlestar galactica

Which measure of central tendency [mean, median, mode] is **most important**? $\emptyset$

Median++… aka Quartiles

Median seems easy: we order the dataset and divide it into two equally sized blocks.

Why two blocks?

Why not four blocks? **Quartiles**

Median++… aka Quartiles

Median seems easy: we order the dataset and divide it into two equally sized blocks.
Why two blocks?
Why not four blocks? **Quartiles**

- **Lower quartile Q1** is the boundary between the lowest 25% of data and the rest.

- **Middle quartile Q2** is the median, i.e. the boundary between top and bottom halves of data.

- **Upper quartile Q3** is the boundary between the highest 25% of data and the rest.

Median++… aka Quartiles

Median seems easy: we order the dataset and divide it into two equally sized blocks.
Why two blocks?
Why not four blocks? **Quartiles**

- **Lower quartile Q1** is the boundary between the lowest 25% of data and the rest.
- **Middle quartile Q2** is the median, i.e. the boundary between top and bottom halves of data.
- **Upper quartile Q3** is the boundary between the highest 25% of data and the rest.

Recipe (easymode): if the number of elements in dataset is divisible by 4, median twice.
Recipe (challengemode): if n is odd, include the median in both halves, and if n is even, split the data in twain. Then, compute the medians for top and bottom halves.

Median++… aka Quartiles

Median seems easy: we order the dataset and divide it into two equally sized blocks.

Why two blocks?

Why not four blocks? **Quartiles**

Quartiles are dividers
They divide my data into 4 pieces

Recipe (easymode): if the number of elements in dataset is divisible by 4, median twice.

Recipe (challengemode): if n is odd, include the median in both halves, and if n is even, split the data in twain. Then, compute the medians for top and bottom halves.

1 2 3 4 5 6 7 8 9 10 11 $n=11$

Example: Compute the quartiles of the data 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

① Median, $Q_2 = 40$

② split & Median

Q_2

6 7 15 36 39 40 | 40 41 42 43 47 49

$Q_1 = 25.5$, $Q_3 = 42.5$

$Q_1 = \frac{15+36}{2} = 25.5$

$Q_3 = \frac{42+43}{2} = 42.5$

Median++… aka Quartiles

Median seems easy: we order the dataset and divide it into two equally sized blocks.

Why two blocks?

Why not four blocks? **Quartiles**

Recipe (easymode): if the number of elements in dataset is divisible by 4, median twice.
Recipe (challengemode): if *n* is odd, include the median in both halves, and if *n* is even, split the data in twain. Then, compute the medians for top and bottom halves.

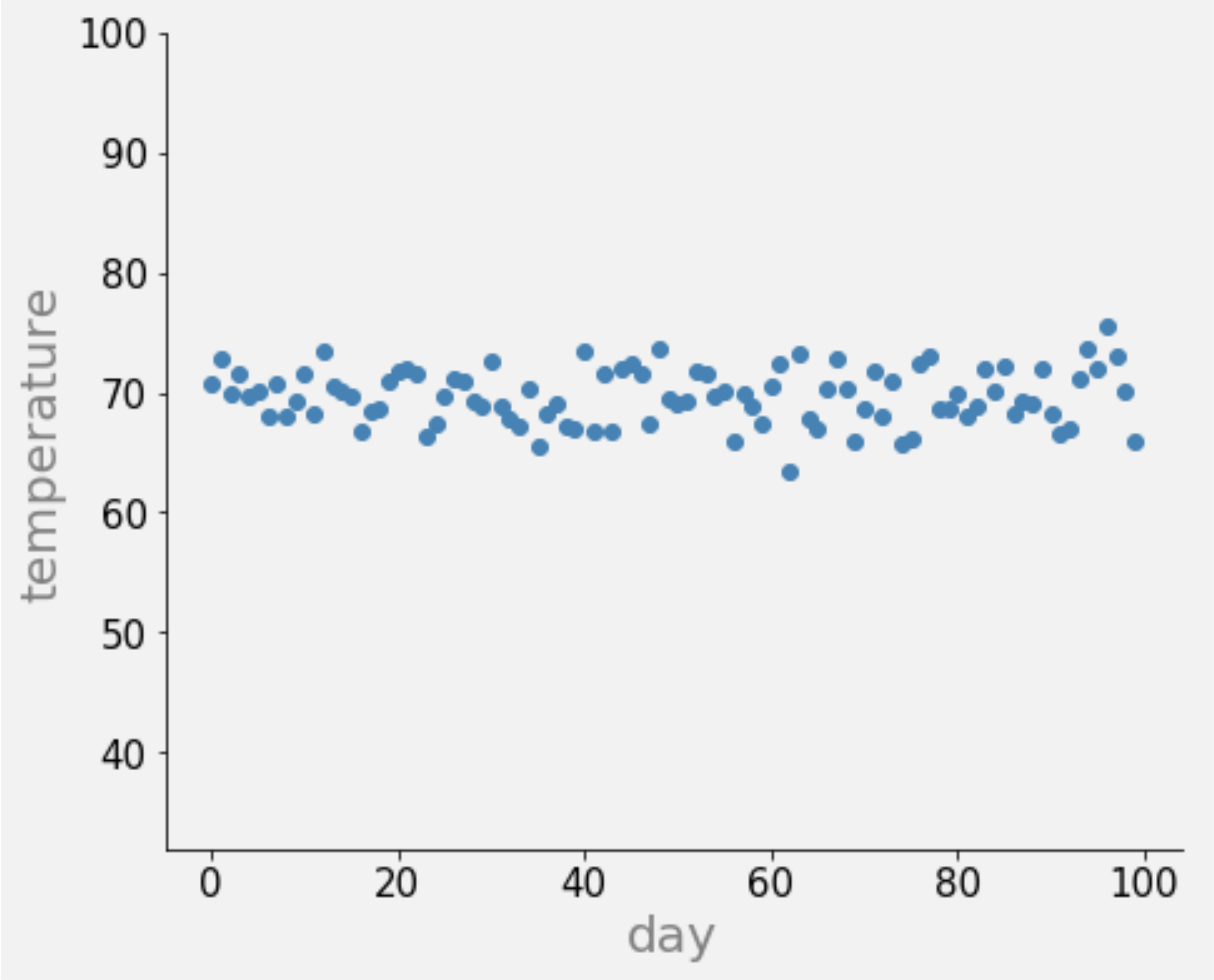
Example: Compute the quartiles of the data 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Bonus: why stop at four? We can define quintiles (split into 5) or the most generic: percentiles. If you're 88th percentile for height, your height is ≥88% of heights.

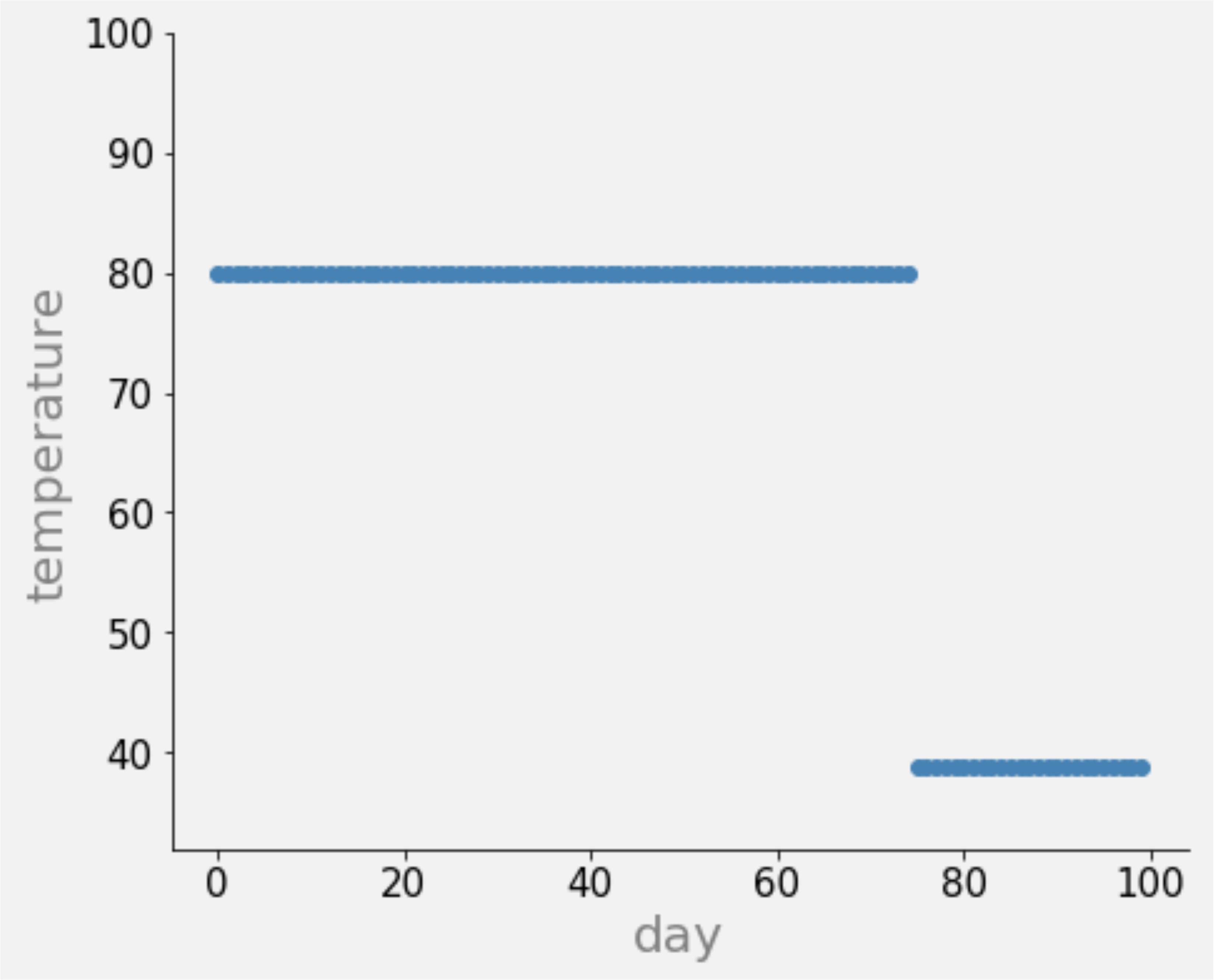
Variability

Mean, median, mode… all measures of centrality. These tell us nothing about the variability or the spread of the data. Sometimes, we may care more about variability than centrality!

Example: A tale of two cities



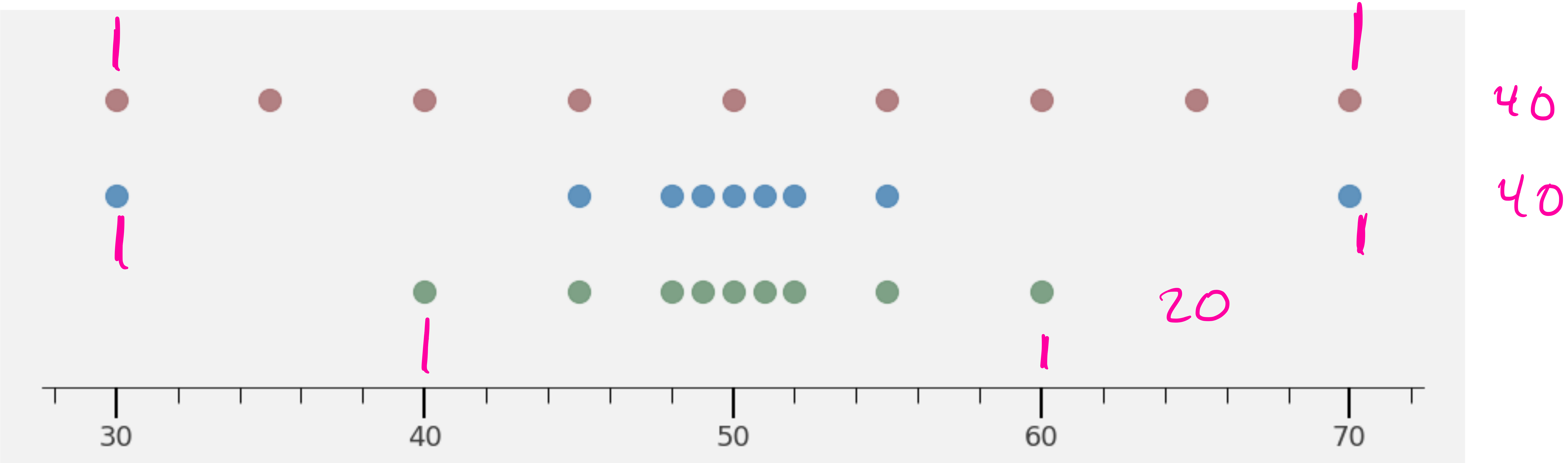
Larremoreville



Ketelsenton

Variability: Range

Definition: Range is the difference between max and min. (Note that this is same as we had in pre-calc, but without the infinities and the open/closed notation.)

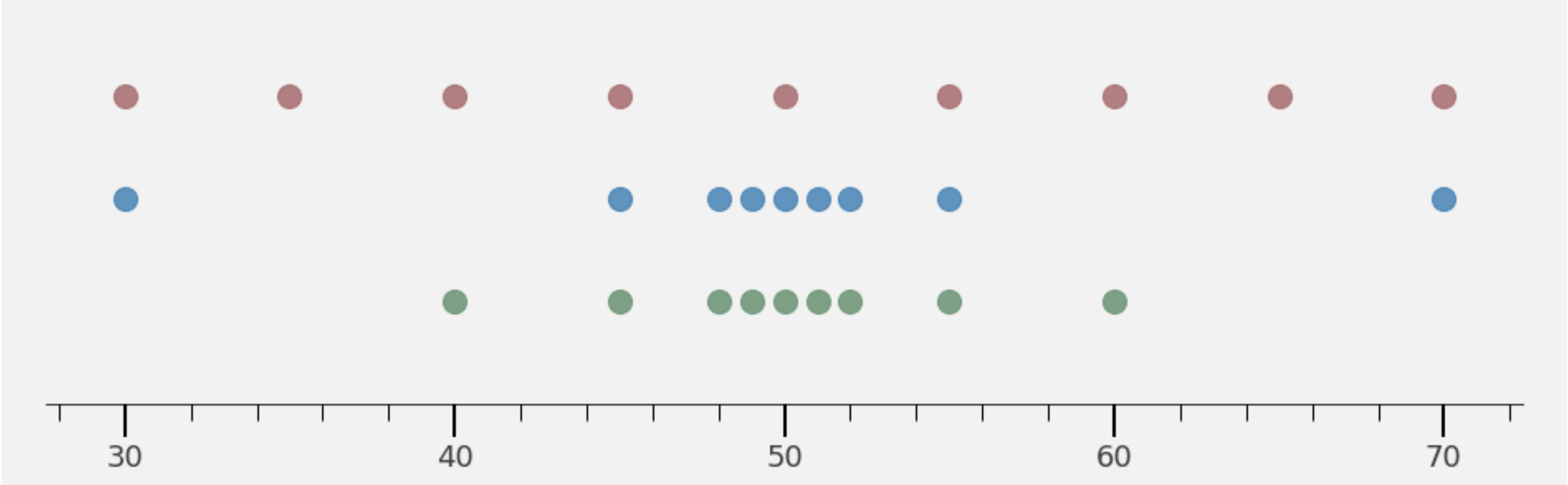


Example: Red, blue, and green all have the same mean. What about the other centralities, median and mode? What about range?

Variability . . .

What do we *see* in the plots here?

Red dots are evenly distributed around the center. Blue and green are uneven. But blue has much bigger deviations. Goal: quantify these difference in a single number.



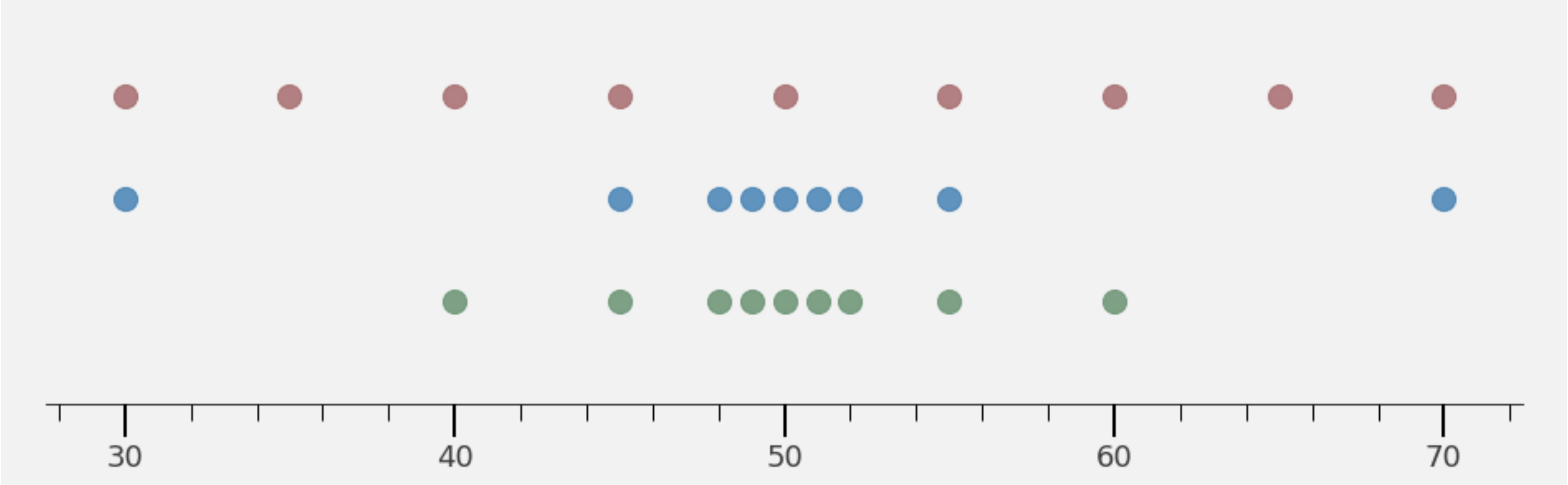
To be robust about the spread around the central value, we could "center" everything:

$$x_1 - \bar{x},\ x_2 - \bar{x},\ x_3 - \bar{x}, \ldots x_n - \bar{x},$$

Variability . . .

What do we *see* in the plots here?

Red dots are evenly distributed around the center. Blue and green are uneven. But blue has much bigger deviations. Goal: quantify these difference in a single number.



To be robust about the spread around the central value, we could "center" everything:

$$x_1 - \bar{x},\ \ x_2 - \bar{x},\ \ x_3 - \bar{x}, \ldots x_n - \bar{x},$$

and then what? Maybe add them up?

$$\frac{1}{n}[x_1 - \bar{x} + x_2 - \bar{x} + x_3 - \bar{x} + \cdots + x_n - \bar{x}] = 0$$

Variance

We're actually going to keep everything positive by squaring the deviation of each point:

$$\frac{1}{n}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x}))^2 + \cdots + (x_n - \bar{x}))^2]$$

And this, my friends, is *almost* the **variance**. Let's write it in a compact form:

almost variance $\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$

sample variance $\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$

Variance

We're actually going to keep everything positive by squaring the deviation of each point:

$$\frac{1}{n}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x}))^2 + \cdots + (x_n - \bar{x}))^2]$$

~~And this, my friends, is *almost* the **variance**. Let's write it in a compact form:~~

The square root of the variance is the **standard deviation**. Sometimes you'll see SD.

$\sqrt{var}$ = std. dev

$\sqrt{\sigma^2} = \sigma$

same units as V. I

Example

Example: Compute the standard deviation of data 2, 4, 3, 5, 6, 4

1 compute mean $\bar{x} = \sum_{k=1}^{n} x_k$ = prev slide = 4

2 compute variance $\text{var} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \cancel{\bar{x}}^{\,4})^2$

$= \frac{1}{6-1}\left[(2-4)^2 + \cancel{(4-4)^2} + (3-4)^2 + (5-4)^2 + (6-4)^2 + \cancel{(4-4)^2}\right]$

$= \frac{1}{5}\left[4 + 1 + 1 + 4\right]$

$= \frac{1}{5}\left[20\right] = 2$

3 $\sqrt{\text{var}}$ = SD

SD $\sigma = \sqrt{2}$