Exploratory Data Analysis
Summary Statistics

Administrivia

- Please activate your Piazza account if you haven't already done so
- No laptops until we get to the in-class notebook part of the lecture
    - [Fried 2006] 64% of students are distracted by other people's laptops
    - [Fried 2006] Second statistically significant distractor: own laptop use
- Be on time and **stay until the end of class**
- If you feel that you would benefit from a smaller classroom environment, consider transferring into Dan's section of the class (Section 002). Only 15 people so far.

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**

Definition: A population is a collection of units (people, songs, tweets, kittens)

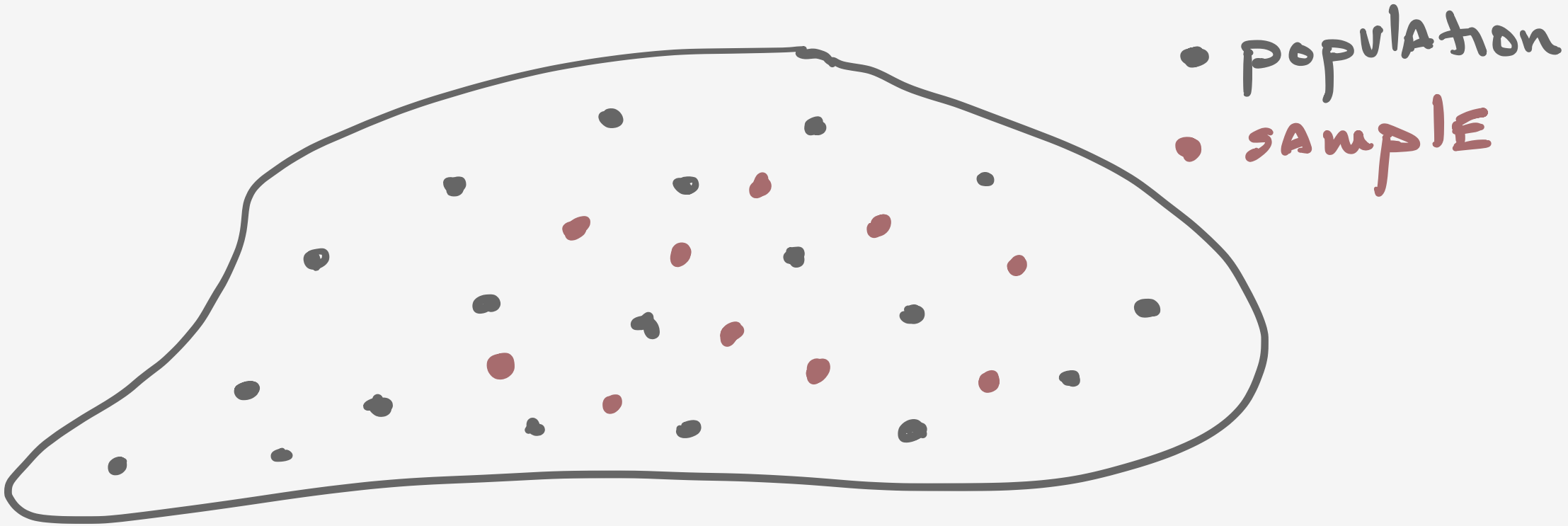
Definition: A sample is a subset of the population

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

- **the population:** DENVER RESIDENTS
- **the sample:** EVERY 50TH PERSON WE PHONE
- **the variable of interest:** HOUSEHOLD

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 50^{th} number on a list of Denver phone numbers between 6pm and 8pm. In this case, what is

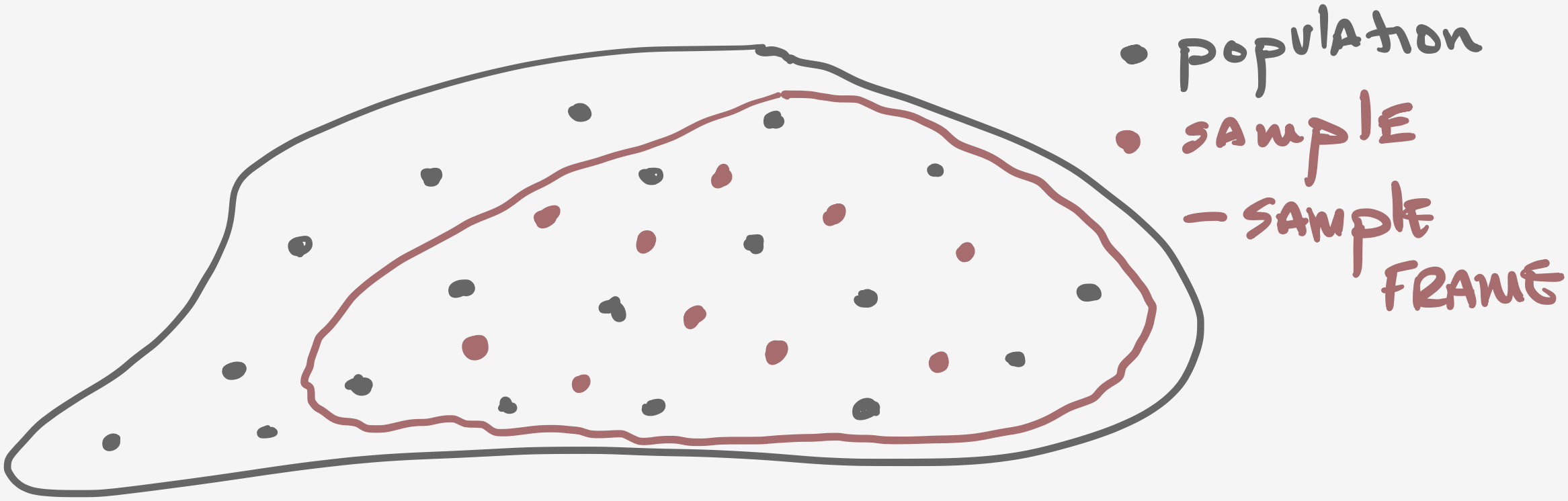
- **the population:**
- **the sample:**
- **the variable of interest:**

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

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**



Samples Types

- **Simple Random Sample**: Randomly select people from sample frame
- **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**: If you have a heterogeneous population that can be broken up into homogeneous groups, randomly sample from each group proportionate to their prevalence in the population

Populations and Samples

Data scientists want to learn about a characteristic in 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 characteristic of a population

Inference!

Exploratory Data Analysis

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

This is useful for summarizing, recognizing patterns, etc. in the data

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

Numerical Summaries

The calculation and interpretation of certain summarizing numbers can help us gain a better understanding of the data.

These sample numerical summaries are called **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 measure of center

- Mean
- Median
- Mode

the Sample Mean

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

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

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

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

$$\bar{x} = \frac{24}{6} = 4$$

$$\Sigma = 2+4+3+5+6+4 = 24$$

the Sample Mean

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

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

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

- sample mean's **advantages**: Easy to calculate,
- sample mean's **disadvantages**: outliers

the Sample Median

Definition: The **sample median** is the "middle" value when the observations are ordered from smallest to largest.

Calculation: Order the n observations from smallest to largest (if there are repeated values, make sure to include each instance of the value).

If n is **odd**: $\tilde{x} = \left(\frac{n+1}{2}\right)^{\text{th}}$ **ordered value**

If n is **even**: $\tilde{x}$ = **the average of** $\left(\frac{n}{2}\right)^{\text{th}}$ **and** $\left(\frac{n+1}{2}\right)^{\text{th}}$ **ordered values**

the Sample Median

Definition: The **sample median** is the "middle" value when the observations are ordered from smallest to largest.

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

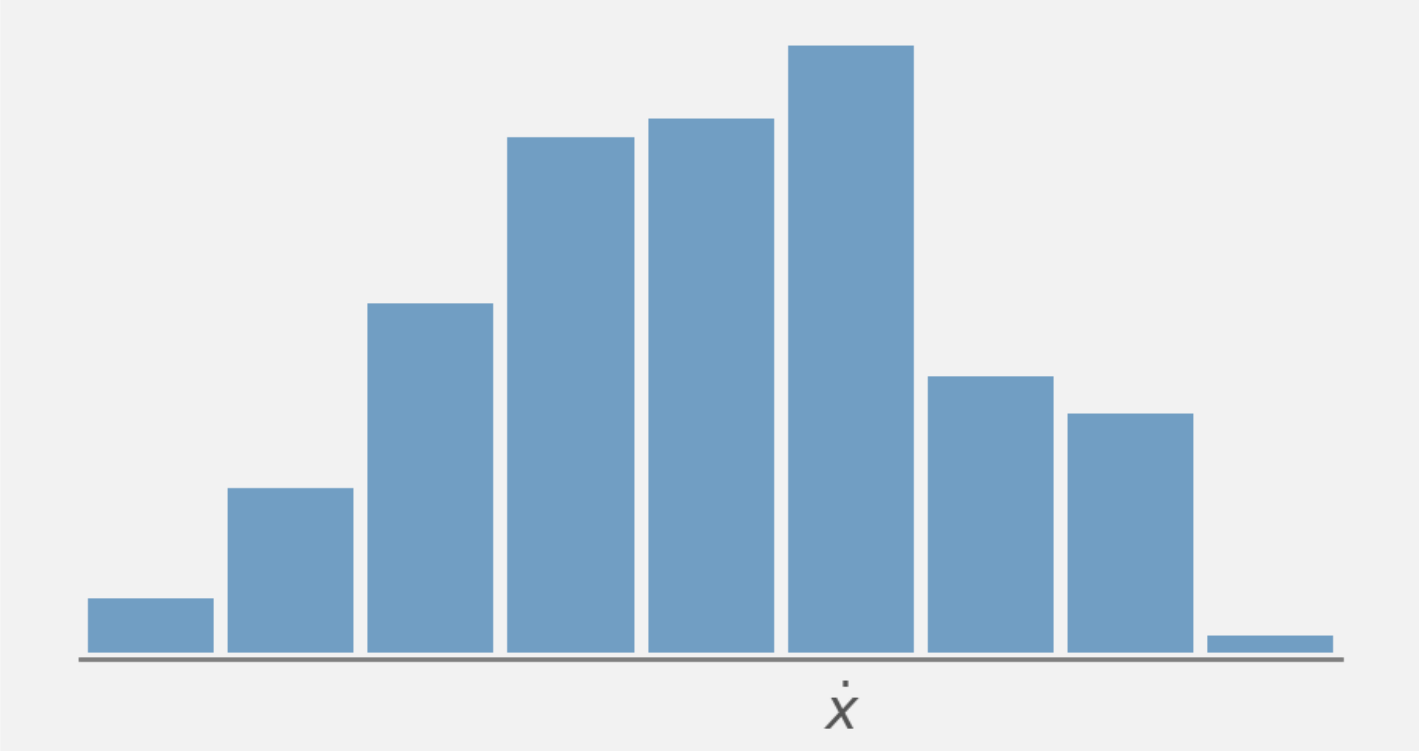
6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

$n = 11$ is ODD

$\tilde{x} = 40$

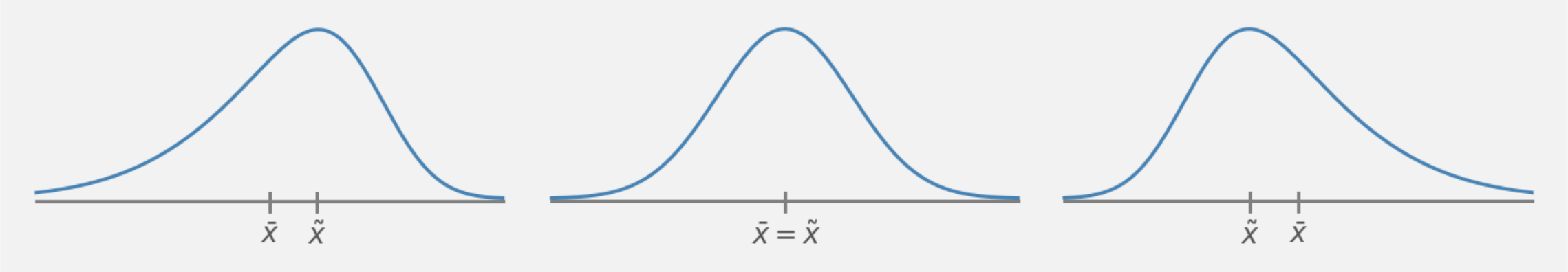
the Sample Mode

Definition: The **sample mode** is simply the value that occurs the most often in the sample



the Mean vs the Median

The population mean and median will not generally be identical. If the population distribution is positively or negatively skewed ...



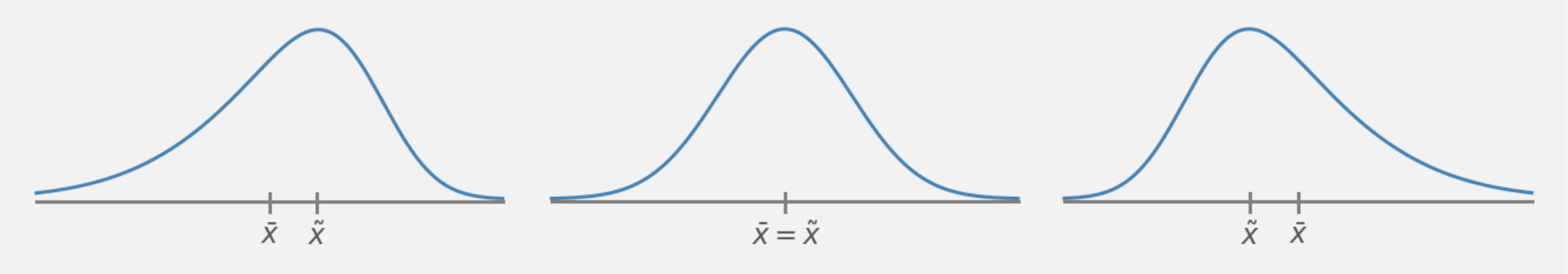
negative skew

symmetric

positive skew

the Mean vs the Median

The population mean and median will not generally be identical. If the population distribution is positively or negatively skewed ...



negative skew

symmetric

positive skew

Which measure of central tendency is the most important?

Other Sample Measures

Quartiles: Divide the data into 4 equal parts.

- Lower quartile Q_1 splits the lowest 25% of the data from the highest 75%
- Middle quartile Q_2 splits the data in half (aka the median)
- Upper quartile Q_3 splits the highest 25% of the data from the lowest 75%

Computation:

1. Use the median to divide the ordered data set into two halves
   - If n is odd include the median in both halves
   - If n is even split the data set exactly in half
2. The lower quartile is median of the lower half. The upper quartile is median of upper half

Other Sample Measures

Quartiles: Divide the data into 4 equal parts.

- Lower quartile Q_1 splits the lowest 25% of the data from the highest 75%
- Middle quartile Q_2 splits the data in half (aka the median)
- Upper quartile Q_3 splits the highest 25% of the data from the lowest 75%

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

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

$Q_1 = \frac{15+36}{2} = 25.5$ $Q_2 = 40$ $Q_3 = \frac{42+43}{2} = 42.5$

Other Sample Measures

Quartiles: Divide the data into 4 equal parts.

- Lower quartile Q_1 splits the lowest 25% of the data from the highest 75%
- Middle quartile Q_2 splits the data in half (aka the median)
- Upper quartile Q_3 splits the highest 25% of the data from the lowest 75%

Can also compute general percentiles, e.g. 37^{th} percentile splits off lower 37% of data

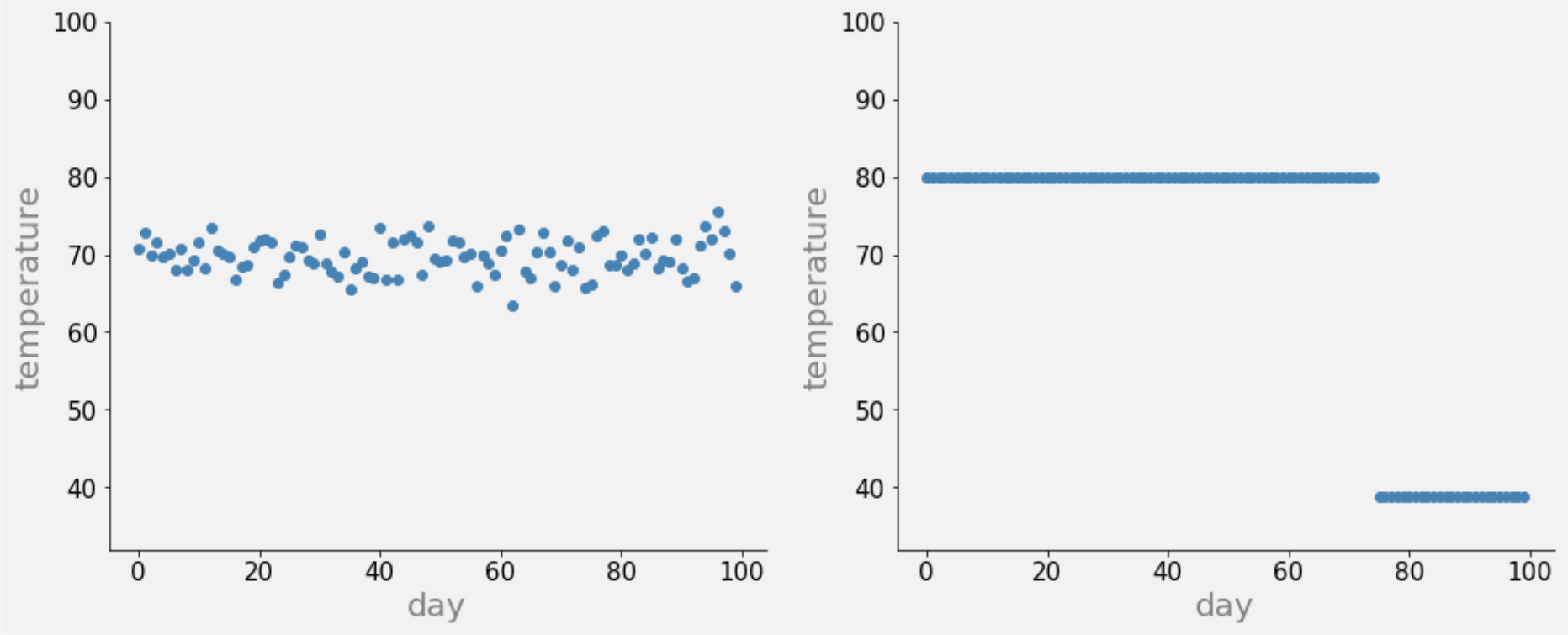
We'll see how to compute these in Python, but won't worry about computation by hand

Variability

So far we've learned about techniques for measuring the center of the data

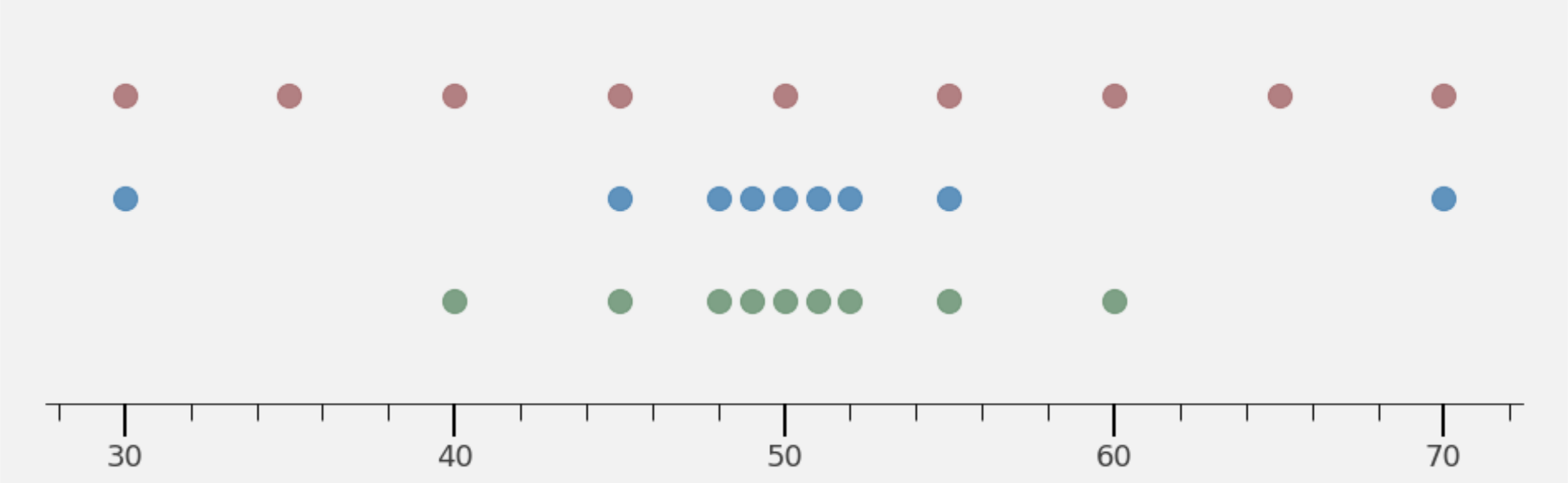
But what about the **spread** of the data?

Example: A Tale of Two Cities



Variability

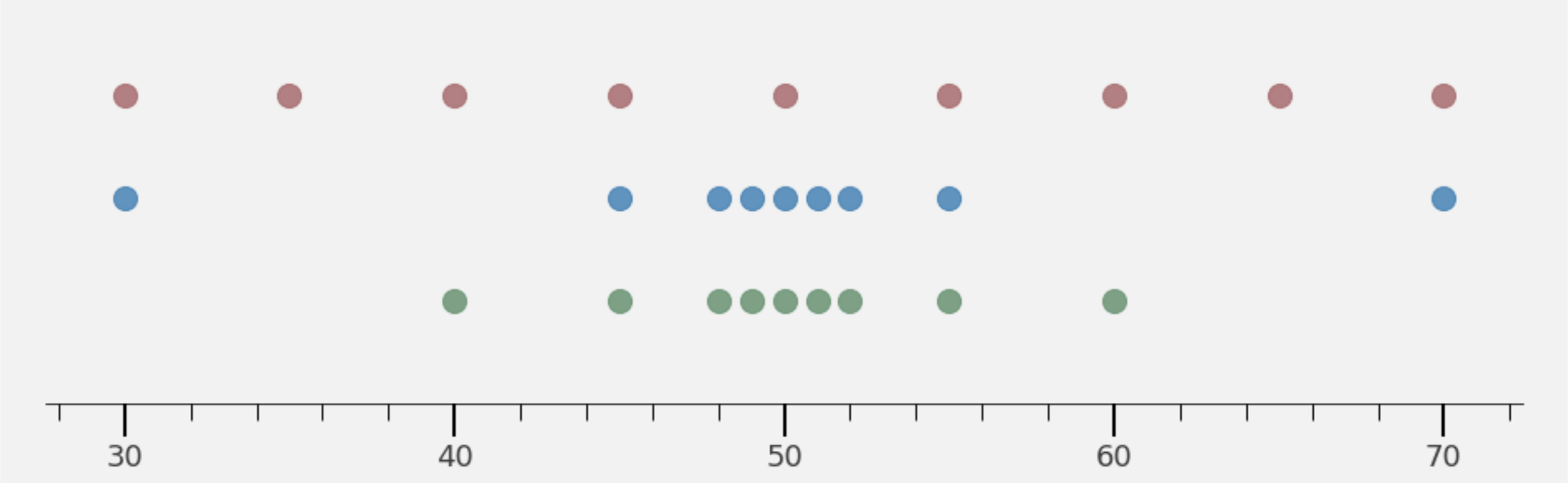
The simplest measure of variability is the **RANGE**



samples with identical measures of centrality but different variability

Variability

The simplest measure of variability is the **RANGE**



samples with identical measures of centrality but different variability

Variability

What if we combined the deviations into a single quantity by finding the average deviation?

A more robust measure of variation takes into account deviations from the mean

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

Variability

What if we combined the deviations into a single quantity by finding the average deviation?

A more robust measure of variation takes into account deviations from the mean

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

So what should we do with these things?

Variability

What if we combined the deviations into a single quantity by finding the average deviation?

A more robust measure of variation takes into account deviations from the mean

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

So what should we do with these things? Add them?

$$\frac{1}{n}\left[(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x})\right]$$

Variability

What if we combined the deviations into a single quantity by finding the average deviation?

A more robust measure of variation takes into account deviations from the mean

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

If we square them first, then it makes all of the deviations positive

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

Variability

The **sample variance**, denoted by s^2, is given by

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

The **sample standard deviation**, denoted by s, is given by the (+ve) square root of the variance

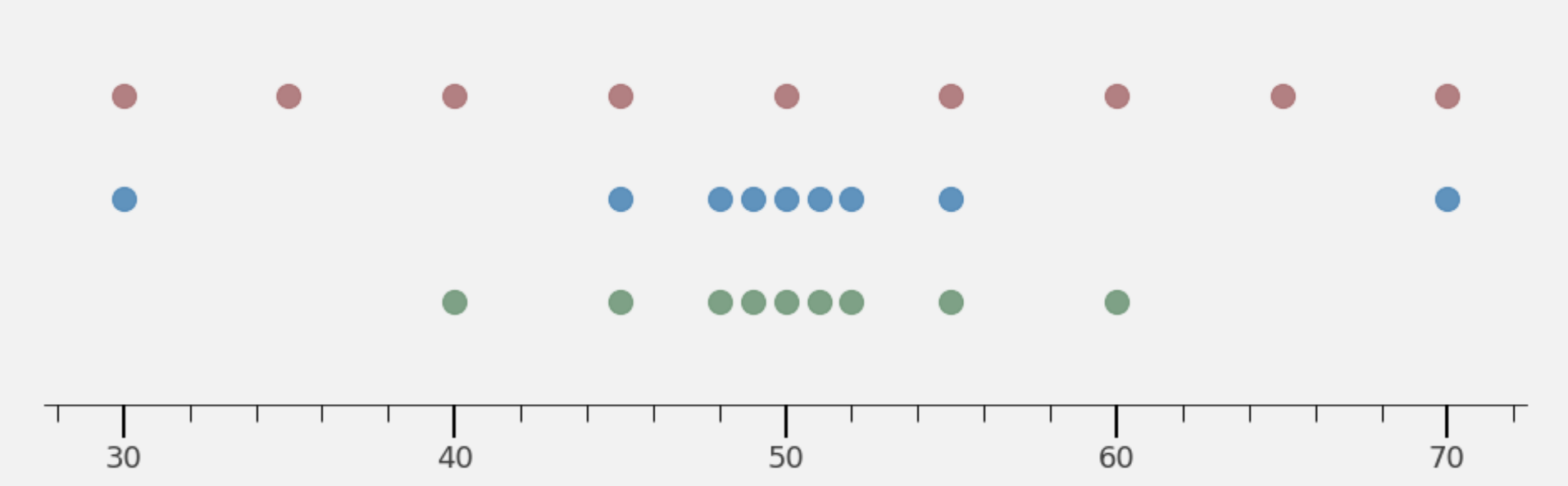
$$s = \sqrt{s^2}$$

Note that the variance and SD are both nonnegative. The units for the SD are the same as data

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

The Interquartile Range

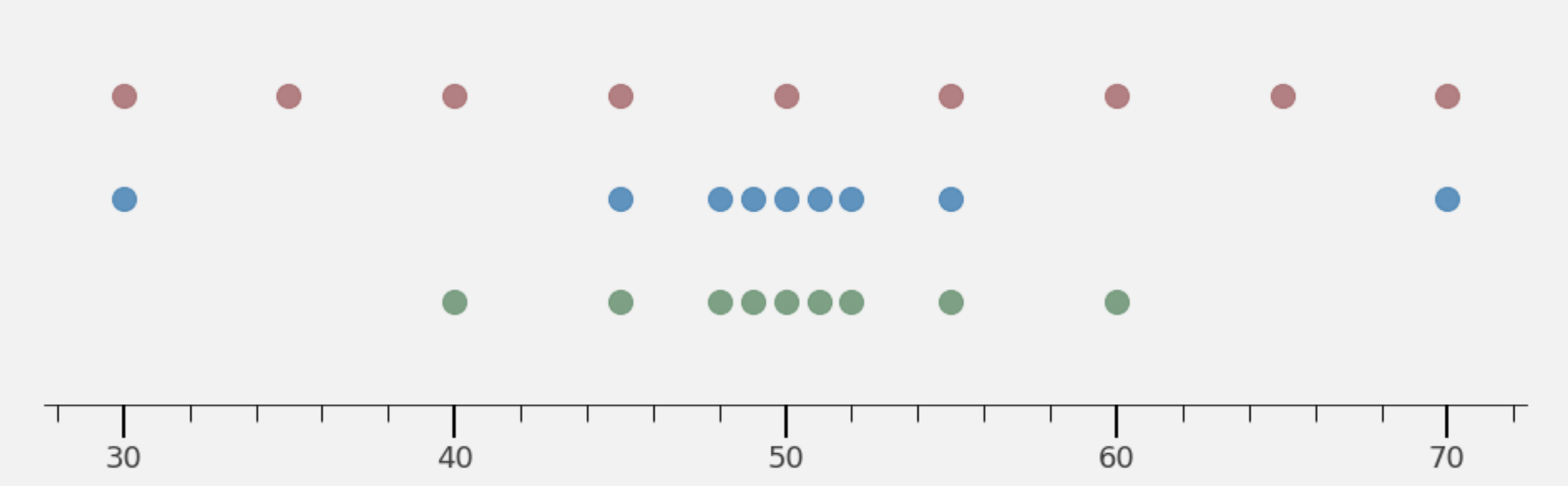
The IQR is defined to be difference between upper and lower quartiles: $IQR = Q_3 - Q_1$



The IQR gives the spread of 50% of the data.

The Interquartile Range

The IQR is defined to be difference between upper and lower quartiles: $IQR = Q_3 - Q_1$



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

Tukey's Five-Number Summary

John Tukey, the father of modern EDA, advocated summarizing data sets with 5 values

min value | lower quartile | median | upper quartile | max value

Tukey's Five-Number Summary

John Tukey, the father of modern EDA, advocated summarizing data sets with 5 values

min value	lower quartile	median	upper quartile	max value
6	25.5	40	42.5	49

Example: The five-number summary of data 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Advantages of the 5-number summary:

- Gives the center of the data
- Gives the spread through the easily computable IQR and range
- Gives an idea of skewness

Tukey's Five-Number Summary

John Tukey, the father of modern EDA, advocated summarizing data sets with 5 values

min value	lower quartile	median	upper quartile	max value
6	25.5	40	42.5	49

Example: The five-number summary of data 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Advantages of the 5-number summary:

- Gives the center of the data
- Gives the spread through the easily computable IQR and range
- Gives an idea of skewness

Next time we'll see how the 5-number summary leads to useful box-and-whisker plots

OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 2 In-Class Notebook

Let's figure out:

- How to compute these summary statistics in Pandas
- How our summary statistics change under transformations of the data