Lecture 2: Exploratory Data Analysis and Summary Statistics

EDA

creative adventures with your data

Announcements and reminders

- **Canvas:** make sure you have looked over the syllabus and schedule

  https://canvas.colorado.edu/courses/24706

- **Piazza:** be on it, because no more emails, and I don't like Canvas very much!

  https://piazza.com/colorado/spring2019/csci3022/

- Get **Jupyter notebook / Anaconda Python** -- make sure you have a working install and check out the Numpy/Pandas tutorial (github/notebooks)

  https://www.anaconda.com/downloads

Populations and samples

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

But, we usually can’t actually see/study the whole population → so we study a **sample**

Populations and samples

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

But, we usually can't actually see/study the whole population → so we study a **sample**

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

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 usually can't actually see/study the whole population → so we study a **sample**

Example: S'pose the city of Boulder wants to estimate its per-household income via a phone survey. They call every 50th number on a list of Boulder phone numbers between 6 PM and 8 PM. In this case, we have:

Population: Boulder residents

Sample: every 50th residents

Variable of Interest: household income

Populations and samples

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

But, we usually can't actually see/study the whole population → so we study a **sample**

Example: S'pose the city of Boulder wants to estimate its per-household income via a phone survey. They call every 50th number on a list of Boulder phone numbers between 6 PM and 8 PM. In this case, we have:

Population:	Boulder residents
Sample:	every 50th person w/ phone who answers
Variable of Interest:	household income

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

Sample 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*/*everything* 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 hope to learn about some **characteristic/variable** of a **population**

But, we usually can't actually see/study the whole population → so we study a **sample**

So how do we make the jump from studying a sample to drawing meaningful conclusions about the characteristic of the population?

Populations and samples

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

But, we usually can’t actually see/study the whole population → so we study a **sample**

So how do we make the jump from studying a sample to drawing meaningful conclusions about the characteristic of the population?

… **inference!**

Exploratory data analysis (EDA)

Before we learn about **inference** though, we first need to learn how to **explore** the data

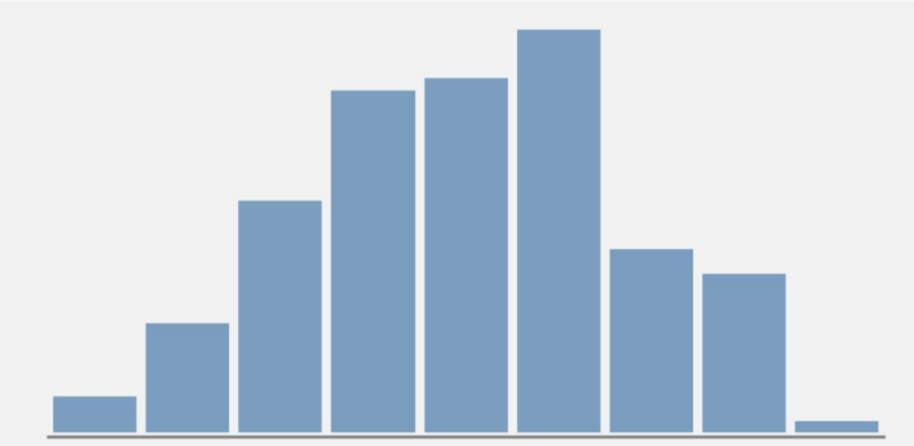
Useful for summarizing, recognizing patterns, etc. in the data

There are two main types 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**



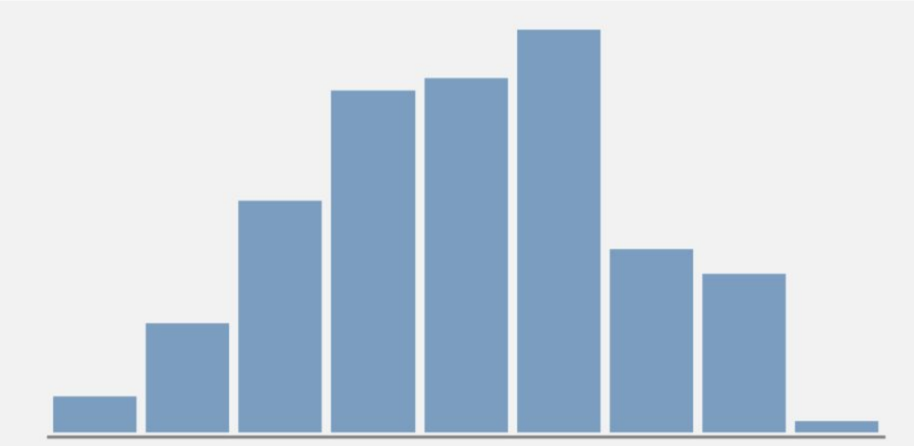
Measures of centrality

Summarizing the “center” -- or better yet -- “**central tendency**” 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

3 main measures:

1) Mean
2) Median
3) Mode



Sample mean

Definition: For a given set of numbers $x_1, x_2, \ldots, x_n$, the **sample mean** is given by:

Also called the **arithmetic average**

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

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

Sample mean

Definition: For a given set of numbers $x_1, x_2, \ldots, x_n$, the **sample mean** is given by:

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

Also called the **arithmetic average**

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

$$\begin{aligned} \bar{x} &= \frac{1}{6}(2+4+3+5+6+4) \\ &= \frac{1}{6} \cdot 24 \\ &= 4 \end{aligned}$$

Sample mean

Definition: For a given set of numbers x_1, x_2, … , x_n, the **sample mean** is given by:

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

Also called the **arithmetic average**

Advantages:

Disadvantages:

Sample mean

Definition: For a given set of numbers x_1, x_2, … , x_n, the **sample mean** is given by:

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

Also called the **arithmetic average**

Advantages: easy to calculate

Disadvantages: outliers can make interpretation misleading

Sample median

Definition: For a given set of numbers 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
- Include multiple instances of repeated values

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

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

Sample median

Definition: For a given set of numbers the **sample median** is the “middle” value when the observations are ordered from smallest to largest.

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

$6, 7, 15, 36, \underline{39, 40}, 41, 42, 47, 49$

$39 + 40 = 79$

$79 \div 2 = \boxed{39.5}$

Sample median

Definition: For a given set of numbers the **sample median** is the "middle" value when the observations are ordered from smallest to largest.

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

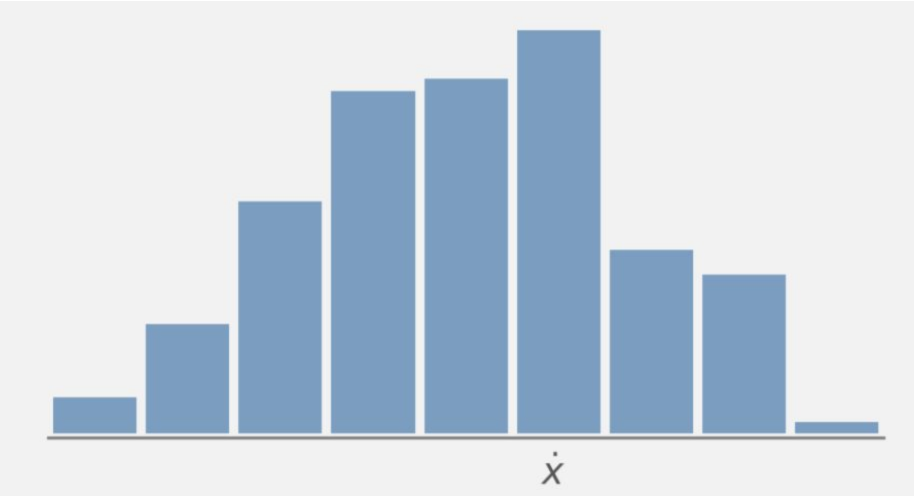
Solution: $n = 10$ is even so it's the average of the middle 2 numbers when sorted:

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

$\rightarrow$ **39.5**

Sample mode

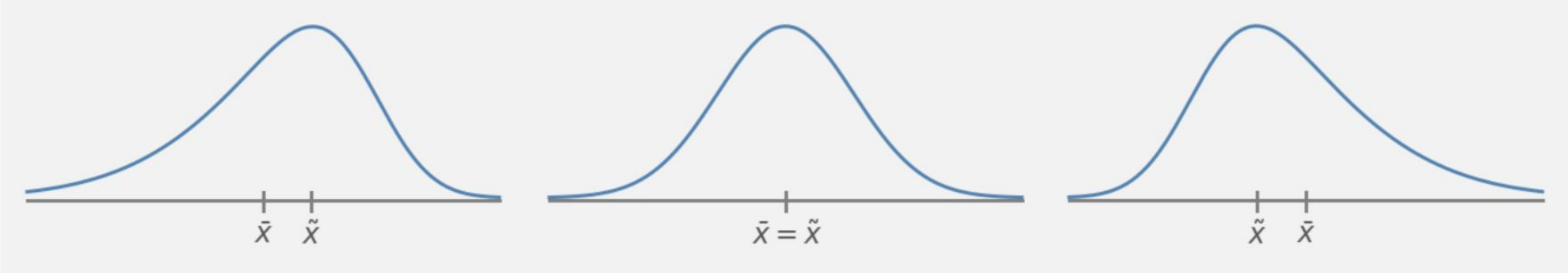
Definition: The **sample mode** is the value that occurs the most often in the sample.



Mean vs median

The population mean and median will generally not be equal.

If the population distribution is positively or negatively skewed …



negative skew
(left-skew)

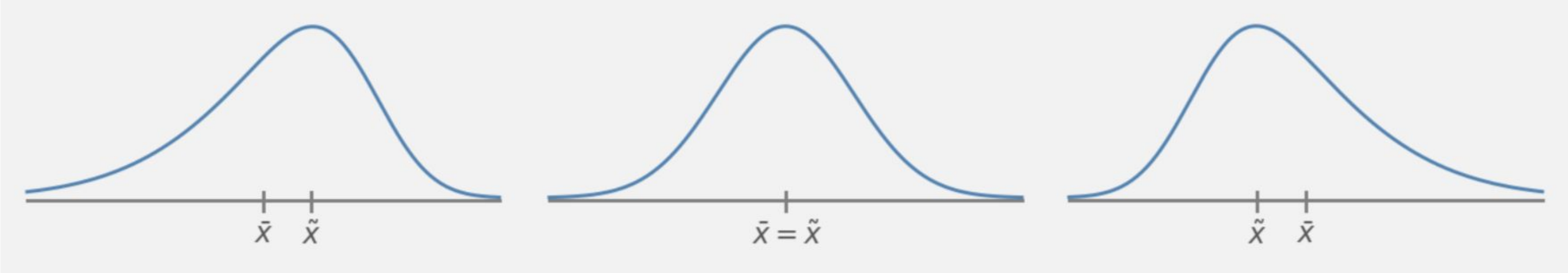
symmetric

positive skew
(right skew)

Mean vs median

The population mean and median will generally not be equal.

If the population distribution is positively or negatively skewed …



negative skew
(left-skew)

symmetric

positive skew
(right skew)

→ Which measure of central tendency is most important?

Other sample measures

Quartiles: Divide the data into 4 equal parts

- Lower quartile (Q_1 or P_{25}) splits the lowest 25% of the data from the other 75%
- Middle quartile (Q_2 or P_{50}) splits the data in half (i.e., the **median**)
- Upper quartile (Q_3 or P_{75}) splits the highest 25% of the data from the lowest 75%

Other sample measures

Quartiles: Divide the data into 4 equal parts

- Lower quartile (Q_1 or P_{25}) splits the lowest 25% of the data from the other 75%
- Middle quartile (Q_2 or P_{50}) splits the data in half (i.e., the **median**)
- Upper quartile (Q_3 or P_{75}) splits the highest 25% of the data from the lowest 75%

Computation:

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

Other sample measures

Quartiles: Divide the data into 4 equal parts

- Lower quartile (Q_1 or P_{25}) splits the lowest 25% of the data from the other 75%
- Middle quartile (Q_2 or P_{50}) splits the data in half (i.e., the **median**)
- Upper quartile (Q_3 or P_{75}) splits the highest 25% of the data from the lowest 75%

Example: Compute the quartiles of the data 6, 7, 15, 36, 39, $\boxed{40,}$ 41, 42, 43, 47, 49

$Q_2 = 40$

$Q_1 = \frac{51}{2} = 25.5$

$Q_3 = \frac{85}{2} = 42.5$

Other sample measures

Quartiles: Divide the data into 4 equal parts

- Lower quartile (Q_1 or P_{25}) splits the lowest 25% of the data from the other 75%
- Middle quartile (Q_2 or P_{50}) splits the data in half (i.e., the **median**)
- Upper quartile (Q_3 or P_{75}) 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

Solution:

1) Data are already sorted
2) Compute median → n=11 is odd, so middle value is median, $Q_2 = 40$
3) Compute Q_1 and Q_3 from first and second halves of data:
   Q_1 = median of first half (6, 7, 15, 36, 39, 40) = (15+36)/2 = 25.5
   Q_3 = median of second half (40, 41, 42, 43, 47, 49) = (42+43)/2 = 42.5

Other sample measures

Quartiles: Divide the data into 4 equal parts

- Lower quartile (Q_1 or P_{25}) splits the lowest 25% of the data from the other 75%
- Middle quartile (Q_2 or P_{50}) splits the data in half (i.e., the **median**)
- Upper quartile (Q_3 or P_{75}) splits the highest 25% of the data from the lowest 75%

Percentiles:

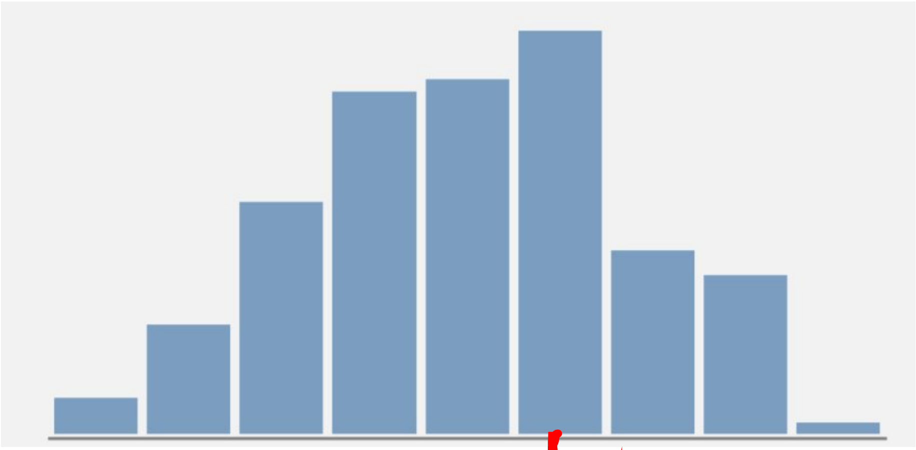
- Generalization of quartiles
- Q_1 is the 25th percentile, P_{25}
- Can also calculate general percentiles:

  e.g., the 16th percentile (P_{16}) splits off the lower 16% of the data.

Variability

So far, we have learned about measuring the **central tendency** of data

But what about the **spread**?

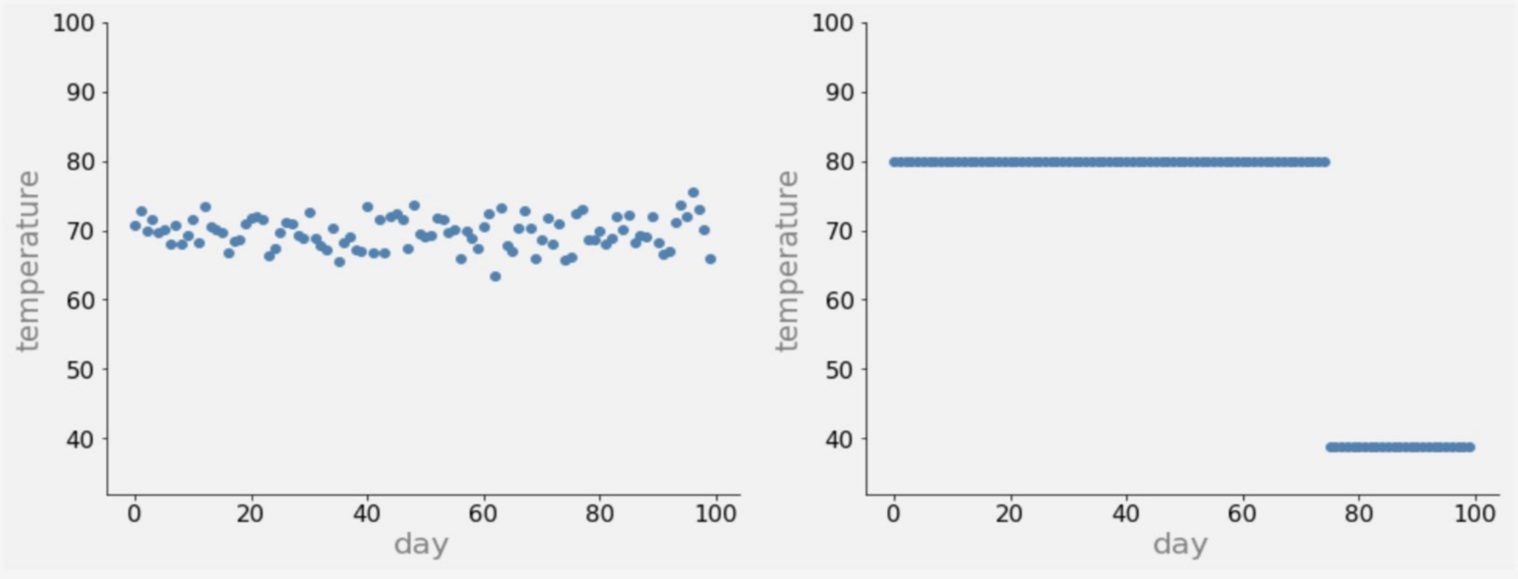


mode

Variability

So far, we have learned about measuring the **central tendency** of data

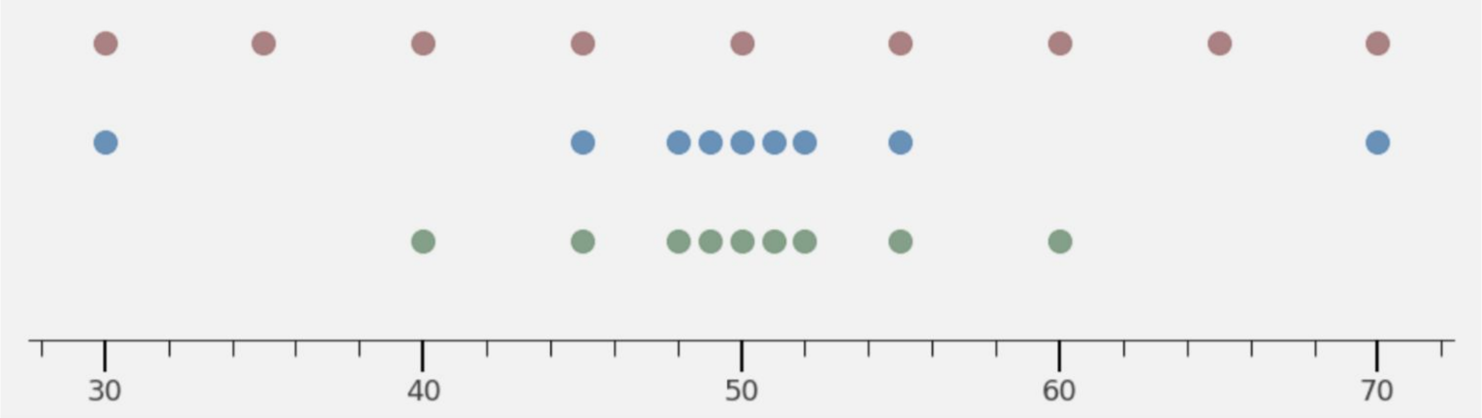
But what about the **spread**?



Variability

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

Definition: The **range** of a sample is the difference between the max and min values



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 do 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 do 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

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

So… what do we do with these things?

… add them?

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

… square, ***then*** add them?

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

Variability

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

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

Definition: The **sample standard deviation**, denoted by s, is given by the (positive) square root of the variance:

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

NB:

- The variance and SD are both **nonnegative** (≥ 0)
- The units for SD are the same as for the data

Variability

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

Variability

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

Solution:

1) Need $\bar{x}$ … = (2+4+3+5+6+4) / 6 = 24 / 6 = 4

2) Calculate s^2 …

$$\begin{aligned} s^2 &= \frac{1}{6-1}\left[(2-4)^2+(4-4)^2+(3-4)^2+(5-4)^2+(6-4)^2(4-4)^2\right] \\ &= \frac{1}{5}\left[4+0+1+1+4+0\right] \\ &= \frac{1}{5}\cdot 10 = 2 \end{aligned}$$

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

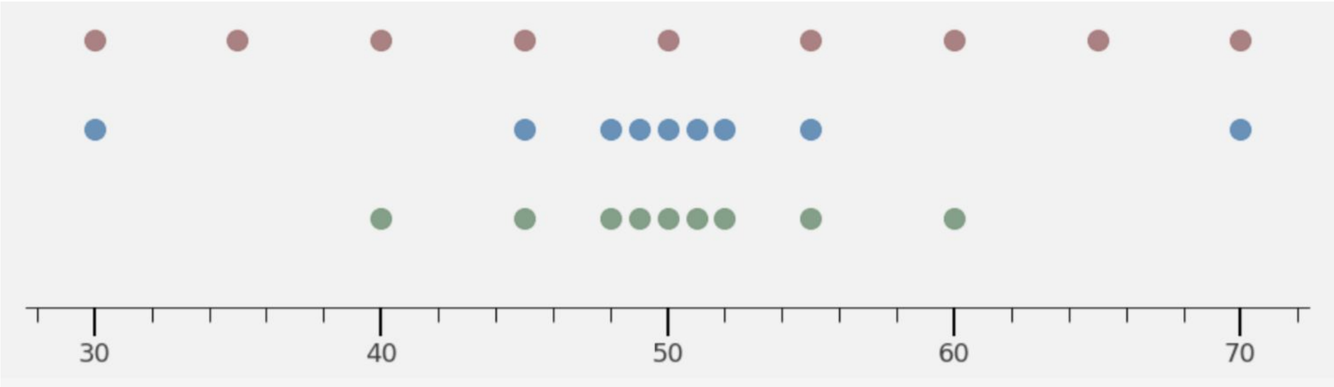
Interquartile range

Definition: The **interquartile range** is defined to be the difference between the upper and lower quartiles:

$$IQR = Q_3 - Q_1$$

→ IQR gives the spread of 50% of the data

Examples:



Interquartile range

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

Interquartile range

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

Solution:

1) Data are already sorted
2) Compute median → n=11 is odd, so middle value is median, Q_2 = 40
3) Compute Q_1 and Q_3 from first and second halves of data:

   Q_1 = median of first half (6, 7, 15, 36, 39, 40) = (15+36)/2 = 25.5

   Q_3 = median of second half (40, 41, 42, 43, 47, 49) = (42+43)/2 = 42.5

4) IQR = Q_3 - Q_1 = 42.5 - 25.5 = 17

Tukey 5-number summary

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

1) Min value
2) Lower quartile
3) Median
4) Upper quartile
5) Max value

Example: Find the 5-number summary of the data 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Tukey 5-number summary

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

1) Min value
2) Lower quartile
3) Median
4) Upper quartile
5) Max value

Example: Find the 5-number summary of the data 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Advantages:

- gives the center of the data
- gives the spread of the data (range in IQR)
- gives and idea of skewness

Tukey 5-number summary

Advantages:

- gives the center of the data
- gives the spread of the data (range in IQR)
- gives and idea of skewness
    - E.g., if Q_2 is closer to Q_1 than to Q_3, then you know the median is “leaning left” (so, distribution is right-skewed)

Next time…

- We’ll see how to visualize this!
 (histograms and box-whisker plots)