

CSCI 3022 Spring 2021

Exploratory Data Analysis

Opening Zoom Poll: which of the following represent violations of the course honor policy?

- ▶ Example 1: For an assignment, Chris searches the internet for relevant codes and copy-pastes them into his Jupyter Notebook. He properly cites the source of the codes.
- ▶ Example 2: For an assignment, Maciej and Felix work together to figure out how to implement the codes, but each works on their own computer and develops their own software.
- ▶ Example 3: For an assignment, Rhonda has a plan for how to implement an algorithm, but isn't sure how to manipulate a Python list in a particular way that she needs to. She searches the internet, finds a fix, and implements it in her code without copying it.

Announcements and To-Dos

Day 1 stuff

Announcements:

1. HW 1 posted later week (due Mon, Feb 1).

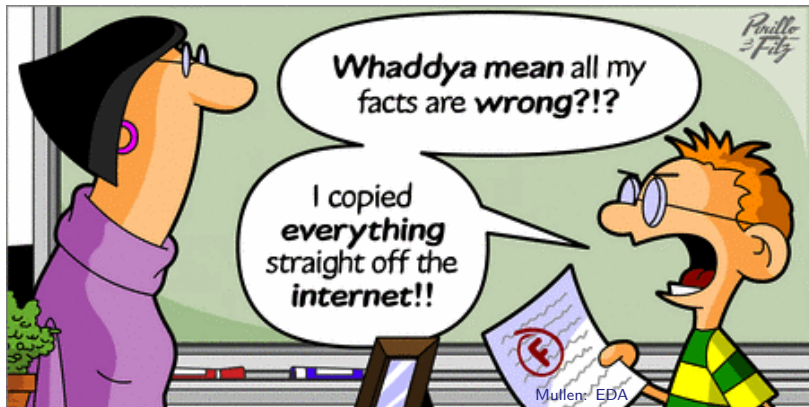
Before next class:

1. Make sure you can access the Canvas page and read the syllabus, get on Piazza
2. Set up some way to back up your work
3. Install Anaconda (or other reliable Jupyter notebook method)
4. Review and complete Numpy/Pandas tutorial
5. Make sure you can load the data in `nb01`

not a homework/turned in

Academic Integrity

1. See the CU Academic Integrity Policy for more details. Here are some highlights.
“Examples of cheating include: copying the work of another student during an examination or other academic exercise (includes computer programming)”
2. “Examples of plagiarism include: . . . copying information from computer-based sources”



Integrity Examples

Example 1: For an assignment, Chris searches the internet for relevant codes and ~~copy-pastes~~ them into his Jupyter Notebook. He properly cites the source of the codes.

code

```
1
2
3
4
5
# You
# key do this
# because they want this
# to compare
# and return
```

you look

fix:

look from code, nice structure/pseudocode

Example 2: For an assignment, Maciej and Felix work together to figure out how to implement the codes, but each works on their own computer and develops their own software.

||

Example 3: For an assignment, Rhonda has a plan for how to implement an algorithm, but isn't sure how to manipulate a Python list in a particular way that she needs to. She searches the internet, finds a fix, and implements it in her code without copying it.

U right process, ... but cite sources

Integrity Examples

Example 1: For an assignment, Chris searches the internet for relevant codes and copy-pastes them into his Jupyter Notebook. He properly cites the source of the codes.

Sol'n: :(. Boo, Chris! Copy-pasting is still not your own work. Correct action: cite a resource, learn the pseudocode or general structure from it and re-create **from the ground up, yourself.**

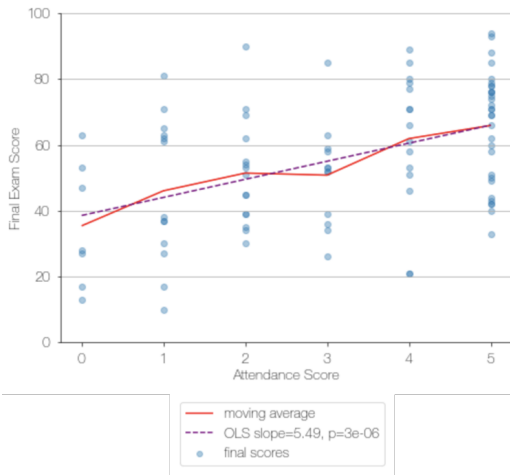
Example 2: For an assignment, Maciej and Felix work together to figure out how to implement the codes, but each works on their own computer and develops their own software.

Sol'n: Awesome! Work together, talk together, develop a theoretical solution together but don't copy the work.

Example 3: For an assignment, Rhonda has a plan for how to implement an algorithm, but isn't sure how to manipulate a Python list in a particular way that she needs to. She searches the internet, finds a fix, and implements it in her code without copying it.

Sol'n: Great!... but just to be safe, cite where you found the fix!

Attend!



Correlation...

is not causation

Try to stay engaged! Take minute papers seriously, and ask questions through any/all mediums available (Zoom, Piazza, minute papers)

The curse of Laptops



“Results showed that students who used laptops in class spent considerable time multitasking and that the laptop use posed a significant distraction to both users and fellow students. Most importantly, the level of laptop use was negatively related to several measures of student learning, including self-reported understanding of course material and overall course performance.”

I know it's a challenge learning remotely! Try to stay focused and hold yourself accountable to a routine with minimal distractions!

<http://www.sciencedirect.com/science/article/pii/S0360131506001436>

Also: <http://journals.sagepub.com/doi/pdf/10.1177/0956797616677314>

And: <http://www.sciencedirect.com/science/article/pii/S0272775716303454>

If at first you don't succeed...

1. When you're asking for help, be sure to explain...
2. what you're trying to do
3. what you think should happen
4. what you get instead (copy/pastes or screenshots work well)
5. what all you have tried
6. if you haven't tried anything, try something first

Learning New Software

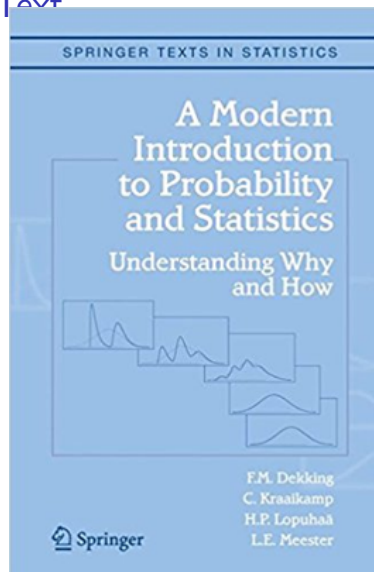
There are 3 major tools to use in learning new software:

1. Pirating similar code found **from course materials**, etc. *# from 1b01...*
2. Official documentation
3. Google searches, often directed to sites like stackexchange. (Don't Copy/Paste! Write from pseudo code, and *cite any sources* if you use them!)

Use (1.) and (2.) often, but be very careful with #3..., and don't hesitate to

1. Ask your instructor or peers for ideas on how to write specific routines, or for their syntax knowledge. Piazza is made for exactly this sort of thing!

Text



A Modern Introduction to Probability and Statistics (MIPS)

by Dekking (et al.)

International, older, and PDF editions will work:
just make sure to match any section numbers that
changed.

Free PDF edition through CU (CU network, or
VPN):

https:

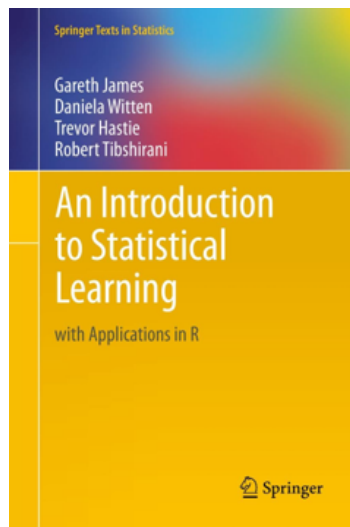
[//www.springer.com/us/book/9781852338961](https://www.springer.com/us/book/9781852338961)

Additional reading will be linked to the course
calendar as needed

Other Texts



Think Stats by Downey (“TS”)



An Introduction to Statistical Learning (“ISL”)

last month
↑

Populations and Samples

Statisticians hope to learn about some characteristic/variable in a population. But we often can't see the whole population; so, we investigate a sample.

Definition: *Population*

A *population* is a collection of units (units can be people, widgets, servings of food, kittens, songs, Tweets, etc.)

Definition: *Sample*

A *sample* is a subset of the population.

Definition: *Variable of Interest (Vol)*

A *characteristic/variable of interest* is something to be measured for each unit.

Populations and Samples

Statisticians hope to learn about some characteristic/variable in a population. But we often can't see the whole population; so, we investigate a sample.

Example: Suppose CU wants to determine the happiness of CS students by a survey.

1 Population (U students?) CSCI students

2 Sample surveys returned

3 Vol happiness → convert to scale/number?

Populations and Samples

Statisticians hope to learn about some characteristic/variable in a population. But we often can't see the whole population; so, we investigate a sample.

Example: Suppose CU wants to determine the happiness of CS students by a survey.

1 *Population*

1a CSCI students, present and future

2 *Sample*

2a 1 in 5 current students polled, less than half respond

3 *Vol*

3a Happiness (a Likert scale?)

Types of Samples

np.random.choice [list]

- ▶ Simple random sample: randomly select people from sample frame *Each and every* person is equally likely to have been selected.
- ▶ Systematic sample: order the sample frame. Choose integer k. Sample every kth unit in the sample frame. *3rd, 8th, 13th, --*
- ▶ 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

Ex: 1000 people
want 10% of them

Sample:

650 men
→ 10% here
65 men

350 women
→
35 women

e.g.
election
polling

Inference and Generalizability

Statisticians learn about a characteristic in a population by studying a **sample**.

A major component of this course is to figure out how they make the jump from sample to population— Statistical Inference!

Statistical inference is can be informally thought of as *the study of missing information*.

Inference and Generalizability

Statisticians learn about a characteristic in a population by studying a **sample**.

A major component of this course is to figure out how they make the jump from sample to population— Statistical Inference!

Statistical inference is can be informally thought of as *the study of missing information*.



Exploratory Data Analysis

Before we learn about inference, we're first going to learn how to explore data. This is helpful for summarizing, recognizing patterns, etc.

There are two main types of explorations: *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. The goal here is to capture something like the “typical” unit with respect to the Vol.

The three most popular measures for centrality

1. The mean
2. The median
3. The mode

The Sample Mean

Definition: Mean

For a given set of n numbers (observations) X_1, X_2, \dots, X_n , the sample *mean* or *arithmetic average* is

$$\frac{\text{add em all up}}{\text{total \# of observations}} \rightarrow \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

"x bar"

1. Advantages:
2. Disadvantages:

The Sample Mean

Definition: *Mean*

For a given set of n numbers (observations) X_1, X_2, \dots, X_n , the sample *mean* or *arithmetic average* is

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

1. Advantages:
2. Disadvantages:

The Sample Mean

' open nbd in it
 ' run first line / open data set
 → put data 'clean_titanic' in some
 di / folder as n.b.

Definition: Mean

For a given set of n numbers (observations) X_1, X_2, \dots, X_n , the sample mean or arithmetic average is

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

add 'em up
 divide by # of #s
 count.

1. Advantages:
 "Easy" to calculate; uses all data;
2. Disadvantages:
 Outliers can matter quite a bit!

The Sample Median

unordered data



Definition: Median

For a given set of n numbers (observations) X_1, X_2, \dots, X_n , the sample median is the middle observation when ordered smallest to largest.

More formally, for data *ordered* smallest to largest $X_{(1)}, X_{(2)}, \dots, X_{(n)}$:

$$\tilde{x} = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ odd} \\ \text{Average of } X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)} & \text{if } n \text{ even} \end{cases}$$

Sorted X

$$X_{(n)} = \min_i X_i$$

"middle two"

1. Advantages

$[0, 1, 2, 3, 1000]$ $[0, 1, 2, 3, 4]$

2. Disadvantages

The Sample Median

Definition: Median

For a given set of n numbers (observations) X_1, X_2, \dots, X_n , the sample *median* is the middle observation when ordered smallest to largest.

More formally, for data *ordered* smallest to largest $X_{(1)}, X_{(2)}, \dots, X_{(n)}$:

$$\tilde{x} = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ odd} \\ \text{Average of } X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)} & \text{if } n \text{ even} \end{cases}$$

1. Advantages

Not using all data makes it less impacted by single observations

2. Disadvantages

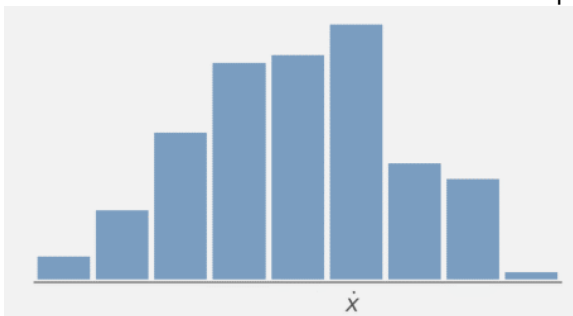
Not using all data makes it less impacted by single observations

The Sample Mode

Definition: *Mode*

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


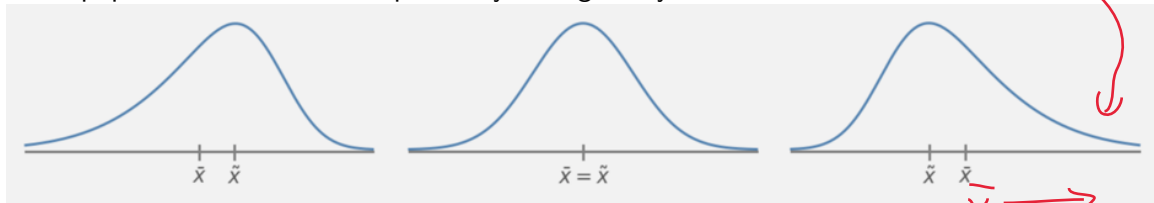
most "typical"?



Skewness: The Mean Versus the Median

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

outlier!
 $(0, 1, 2, 3, 1000)$

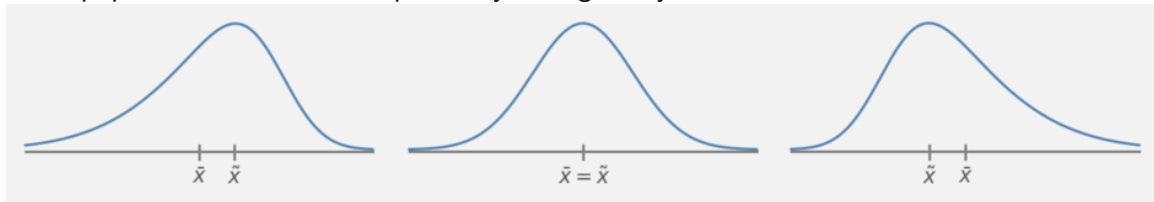
Mean < Median
“Left skew”

Mean \approx Median
“Symmetric”

Mean > Median
“Right skew”

Skewness: The Mean Versus the Median

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



Mean < Median
“Left skew”

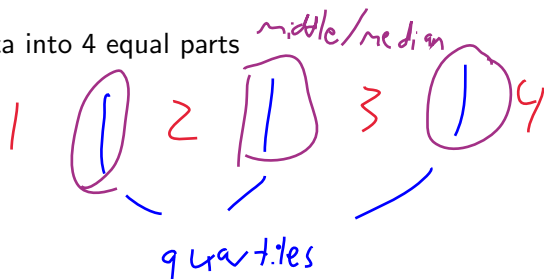
Mean \approx Median
“Symmetric”

Mean > Median
“Right skew”

We’ll talk more about these types of pictures next time, but in brief: *skewness* tracks **where** points that are “far from center” lie. For example, the data set $[1, 2, 1, 1, 1, 2, 100]$ is right-skewed, since the only “far away” point is to the positive/right side of the data.

Quartiles

Definition: *Quartiles* Divide the data into 4 equal parts



Quartiles

Definition: *Quartiles* Divide the data into 4 equal parts

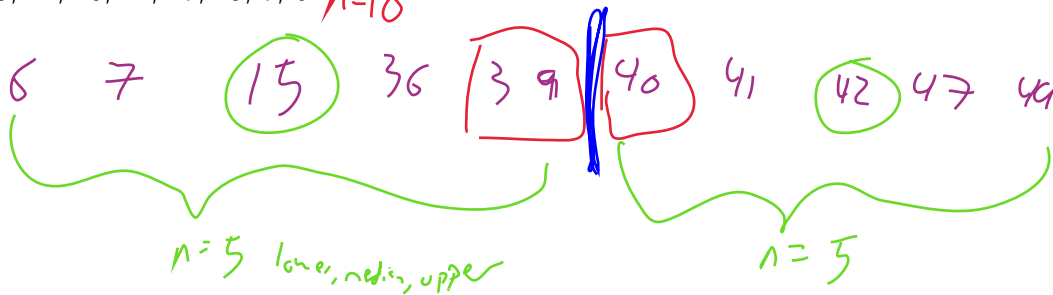
Example: Calculations of the median and quartiles:

Calculate the sample median and quartiles of the data:

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

$n=10$

Median = 39.5



Quartiles: $[15, 39.5, 42]$

Quartiles and Percentiles are generalizations of quartiles.

Quartiles

Definition: *Quartiles* Divide the data into 4 equal parts

Example: Calculations of the median and quartiles:

Calculate the sample median and quartiles of the data:

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

First, sort the data: 6, 7, 15, 36, 39, 40, 41, 42, 47, 49

Quantiles and *Percentiles* are generalizations of quartiles.

Quartiles

Definition: *Quartiles* Divide the data into 4 equal parts

Example: Calculations of the median and quartiles:

Calculate the sample median and quartiles of the data:

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

First, sort the data: 6, 7, 15, 36, 39, 40, 41, 42, 47, 49

Now chop it up!: 6, 7, 15, 36, 39, *MIDDLE*, 40, 41, 42, 47, 49

Quantiles and *Percentiles* are generalizations of quartiles.

Quartiles

Definition: *Quartiles* Divide the data into 4 equal parts

Example: Calculations of the median and quartiles:

Calculate the sample median and quartiles of the data:

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

First, sort the data: 6, 7, 15, 36, 39, 40, 41, 42, 47, 49

Now chop it up!: 6, 7, 15, 36, 39, *MIDDLE*, 40, 41, 42, 47, 49

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

15 is the first quartile

39.5 is the median or second quartile

42 is the third quartile

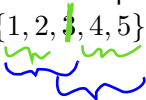
Quantiles and *Percentiles* are generalizations of quartiles.

Quartiles

There are multiple ways to define a quartile! Suppose we have a data set that contains: $\vec{X} = \{1, 2, 3, 4, 5\}$. What are its quartiles?

Quartiles

There are multiple ways to define a quartile! Suppose we have a data set that contains:
 $\vec{X} = \{1, 2, 3, 4, 5\}$. What are its quartiles?

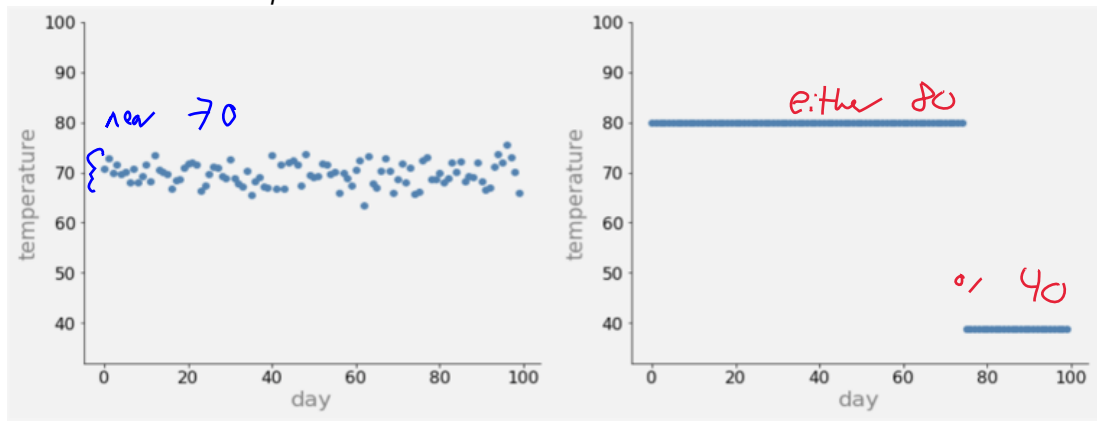


1. This depends on whether or not we include the median (3) in each upper and lower halves. If we do, we get 2 and 4. If we don't, we get 1.5 and 4.5.
2. There is not universal agreement between statistician *nor* software packages over which to use
3. It turns out that there's two other methods that *interpolate* the data: the median might be 1/4 of the way between two observations (1.75 and 4.25) *or* it might sit somewhere fractionally between e.g the 5/21th point and the 6/21th point. Ugh!
4. **Python** includes the median in each smaller "half." So for our purposes, this set has quartiles of 2 and 4.

Dispersion and Spread

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

But what about the *spread*?



Left: San Francisco

Right: Mullensville

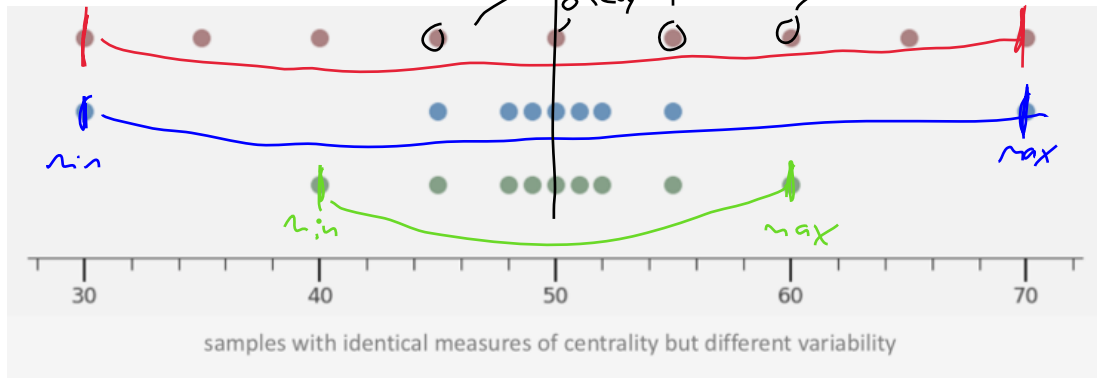
The Range

Simplest measure of variability: The range.

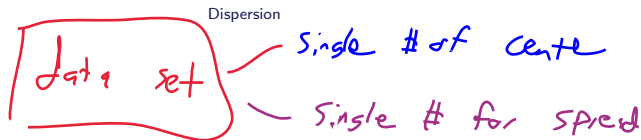
$\bar{x} = 50$

0 way 5 way 10 way

min max

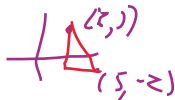


Deviation



We probably care about how far away points are from their average. "Far," of course, is actually a math word.

- ▶ The distance between two numbers a and b is $D = |a - b|$.
- ▶ The distance between two points (a_1, a_2) and (b_1, b_2) is $D = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$



We want to use the distance from the mean. But which type distance? Squared or not?

!!

Deviation

We probably care about how far away points are from their average. “Far,” of course, is actually a math word.

- ▶ The distance between two numbers a and b is $D = |a - b|$.
- ▶ The distance between two points (a_1, a_2) and (b_1, b_2) is $D = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$

We want to use the distance *from the mean*. But which type distance? Squared or not? For each datum X_i , the *deviation from the mean* of X_i is

“absolute deviation”

$$|X_i - \bar{X}|$$

want
average
distance

Variance and Standard Deviation

Definition: Sample Variance

The sample variance, denoted by S^2 , is given by:

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

dist. from mean

"average" squared distance from the mean

"average, but shifted"

Do NOT USE $\text{np.var}(X)$.

The sample standard deviation, denoted by S , is the (positive) square root of the variance:

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

divide by n , not $n-1$

$(X_i - \bar{X})$ units of data.

Note that S^2 and S are both nonnegative. The unit for S is the same as the unit for each of the X_i .

Variance and Standard Deviation

Definition: *Sample Variance*

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

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

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

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

Note that s^2 and s are both nonnegative. The unit for s is the same as the unit for each of the X_i .

Standard Deviation

Example: Calculation of the SD

Data (units in dollars): 2,4,3,5,6,4.

$$\begin{array}{l}
 \text{Variance: } (2 - \bar{x})^2 + (4 - \bar{x})^2 + (3 - \bar{x})^2 \\
 + (5 - \bar{x})^2 + (6 - \bar{x})^2 + (4 - \bar{x})^2 \\
 \hline
 6 - 1
 \end{array}$$

Standard Deviation

Example: *Calculation of the SD*

Data (units in dollars): 2,4,3,5,6,4.

Since we mean business, we need the average first.

$$\bar{X} = \frac{2 + 4 + 3 + 5 + 6 + 4}{6} = \frac{24}{6} = 4$$

Now let's compute the deviations...

vectorized deviations

$$\overbrace{[(X - \bar{X})^2]}^{\text{vectorized deviations}} = [(2 - 4)^2, (4 - 4)^2, (3 - 4)^2, (5 - 4)^2, (6 - 4)^2, (4 - 4)^2]$$

and sum and “average” those!

$$s^2 = \frac{4 + 0 + 1 + 1 + 4 + 0}{6} = 2$$

The Interquartile Range

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

$$IQR = Q_3 - Q_1$$

It's a spread measure standardly used in *box plots*, which we introduce formally next time.

Tukey's Five 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

Advantages:

- gives the center of the data
- gives the spread of the data (range in IQR)
- gives an idea of skewness (compare how far away Q1 and Q3 are from median!)

Next Lecture: Visual EDA!

Collapsing our data into a few descriptive numbers is pretty valuable!

...but *summary statistics* invariably throw away a lot of detail and nuance. Maybe we should consider visualizing the data to include more information?