8-28 EDA

Wednesday, August 28, 2019 12:54 PM

8-28 EDA

8-28: Exploratory Data Analysis (EDA)
CSCI 4022 Fall 19

EDA Fall 2019 1 / 22

Today:
100% less
powerpoint.

Announcement and Reminders

- Canvas has a quizlet!
- Sign up for course Piazza
- CA/graduate office hours are in ECAE190-191
- Get Jupyter notebook/Anaconda environment running
- Do nb00, pandas and Numpy tutorials.

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 (VoI)*
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* CS students (present/future).
2. *Sample* $4/100 \sim 6/100$.
3. *VoI* "Happiness"

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 *VoI*

3a Happiness (a Likert scale?)

Types of Samples

- 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.
- 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

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 VoI.

if I found a "new" observation, what does that look like?

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) $X_1, X_2, \ldots, X_n$, the sample *mean* or *arithmetic average* is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + \cdots X_n}{n}$$

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

c

1. Advantages: Everything is weighted equally; easy to compute.
2. Disadvantages:

outliers

Ex: $X = [2, 3, 3, 4, 100]$. $\bar{X} = \frac{102}{5} = 2\cancel{0}2$.

The Sample Mean

Definition: *Mean*
For a given set of n numbers (observations) $X_1, X_2, \ldots, X_n$, the sample *mean* or *arithmetic average* is

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

1. Advantages:

2. Disadvantages:

The Sample Mean

Definition: *Mean*
For a given set of n numbers (observations) $X_1, X_2, \ldots, X_n$, the sample *mean* or *arithmetic average* is

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

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

The Sample Median

Definition: *Median*
For a given set of n numbers (observations) $X_1, X_2, \ldots, X_n$, the sample *median* is the middle observation when ordered smalled to largest.
More formally, for data *ordered* smallest to largest $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$:

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

1. Advantages
   Robust to outliers. minimizes
2. Disadvantages
   Doesn't use all the data

function total distance from ith observation to the number c.

$$f(c) = \sum_{i=1}^{n} |X_i - c|$$

The Sample Median

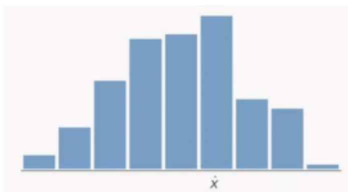
Definition: *Median*
For a given set of n numbers (observations) $X_1, X_2, \ldots, X_n$, the sample *median* is the middle observation when ordered smalled to largest.
More formally, for data *ordered* smallest to largest
$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$:

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

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

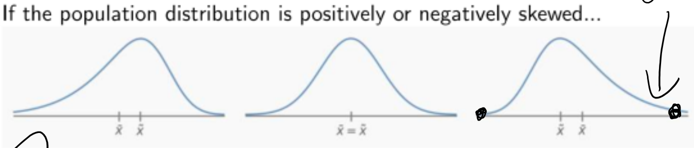
Definition: *Mode*
The sample *mode* is the value that occurs the most often in the sample.



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...

outliers exist in the +X



Mean < Median	Mean ≈ Median	Mean > Median
"Left skew"	"Symmetric"	"Right skew"

outliers exist in the −X direction.

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

1) Sort: 6, 7, $\boxed{15}$, 36, 39, | 40, 41, $\boxed{42}$ 47, 49

middle
= 39.5

Quantiles and *Percentiles* are generalizations of quartiles.

Quartiles

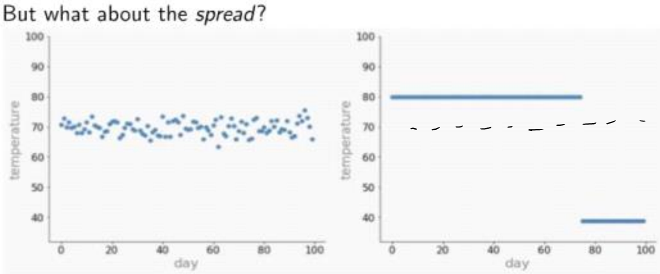
Definition: *Quartiles* Divide the data into 4 equal parts
Example: Calculations of the median and quartiles:
Calculate the sample median and quartiles of the data:
36, 15, 39, 41, 40, 42, 47, 49, 7, 6

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

Quantiles and *Percentiles* are generalizations of quartiles.

Dispersion and Spread

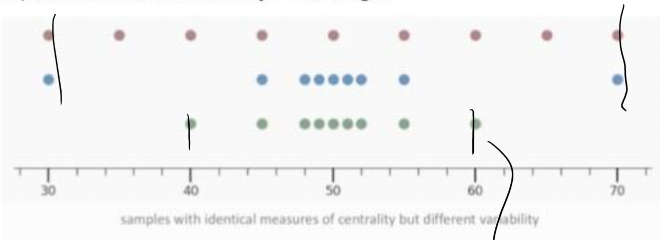
So far, we have learned about measuring the central tendency of data
But what about the *spread*?



Left: San FranZachsco
Right: Mullensville

The Range

Simplest measure of variability: The range. → range = 40



range = 20

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

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

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.

Variance and Standard Deviation

need

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

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

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.

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

then: $(X_i - \bar{X}) = [(2-4), (4-4), (3-4), (5-4), (6-4), (4-4)]^T$

$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{2^2 + 0^2 + 1^2 + 1^2 + 2^2 + 0^2}{5} = \frac{10}{5} = 2$

$s = \sqrt{2}$.

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.

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

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...

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

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...

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

and sum and "average" those!

$$s^2 = \frac{4+0+1+1+4+0}{5} = 2$$

For a standard deviation of $s = \sqrt{2}$.

The Interquartile Range

(nice w/ outlies)

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 Time: 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?