Calculus-based Intro to Probability & Statistics

Math 11: Discussion 1

TA: Thu Nguyen
Tue. 03/31/2020

Recap of discussion 1:

1. Assignments are spread out across the week, each at different times.
2. Some examples of what you will be able to do after the course.
3. Examples of how to describe a data set, and how the metrics change given small changes in the data set.

Contents

1.1 Logistics + upcoming deadlines

Logistics:

- All meetings (discussions/office hours) are on Zoom and not recorded (for privacy and legal concerns).
- Office hours: Wednesdays at 5.00 – 7.00 pm.
- Email: thn003@ucsd.edu (preferred over messages on Canvas).

Upcoming deadlines.

Assignments	Deadlines
Homework - Ch. 2	Mon. 04/06
Quiz - Ch. 2	Tue. 04/07
Lab 1	Mon. 04/06

Note: Assignments are spread out across the week.

1.2 Motivation for the course

- PROBABILITY: to model uncertainty/chances and to possibly predict the future.

1. If we roll 2 die and sum up the 2 numbers, which is more likely: an even or odd number?

> Odd and even are equally likely.
>
> *Reason.*
> - For each dice, equal chances of an odd or even number.
> - Given 2 die, 4 possible combinations of *odd* and *even*.
> - Let S be the sum:
>
> $$S \text{ is } Odd \text{ if } S = Odd + Even = Even + Odd$$
> $$S \text{ is } Even \text{ if } S = Even + Even = Odd + Odd$$

2. Suppose we draw 2 cards, rank the following events from most likely:
(A) a pair of *Aces*, (B) 2 black *Aces* vs. (C) a black *Ace* and a black 10?

> From most likely: (A) 2 *Aces*, (C) a black *Ace* and a black 10, (B) 2 black *Aces*
>
> *Reason.* We calculate the probability of successfully getting the 1^{st} and then the 2^{nd} cards:
>
> $$\mathbb{P}(A) = \frac{4}{52} \cdot \frac{3}{51}; \quad \mathbb{P}(B) = \frac{2}{52} \cdot \frac{1}{51}; \quad \mathbb{P}(C) = \frac{4}{52} \cdot \frac{2}{51};$$

- STATISTICS: to analyze the past and to possibly interpolate to make inferences about the future.

1. How do we compare Math 11 grades across offerings, say between Fall '19 and Winter '20?

> We compare across different metrics:
>
> $$\underbrace{\text{min}, \quad \text{max}, \quad \text{mean}, \quad \text{median},}_{\text{critical data points}} \quad \underbrace{\text{standard deviation}, \quad IQR}_{\text{spread of data}}$$
>
> Note that depending on the data, some metrics are more meaningful than the others:
>
> $$\underbrace{\text{(mean, standard deviation)}}_{\text{more normally distributed / spread-out data}} \qquad vs. \qquad \underbrace{\text{(median, IQR)}}_{\text{more extreme / skewed data}}$$

2. On CAPES, there are records of the number of hours spent and grades, how can we make use of that?

> Suppose we believe that the more we study, the better the grades.
>
> In particular, every 1 additional hours spent would increase the raw score by 5%, then:
>
> $$\underbrace{\text{raw score} = 50\% + 5 * (\text{hours spent}) + \text{noise}}_{\text{Linear regression model}}$$

1.3 Lecture 1 review

Recall that given some data points, we can describe with a number of metrics:

$$\underbrace{\text{min}, \quad \text{max}, \quad \text{mean}, \quad \text{median},}_{\text{critical data points}} \quad \underbrace{\text{standard deviation}, \quad IQR}_{\text{spread of data}}$$

Note that depending on the data, some metrics are more meaningful than the others:

$$\underbrace{\text{(mean, standard deviation)}}_{\text{more normally distributed / spread-out data}} \qquad vs. \qquad \underbrace{\text{(median, IQR)}}_{\text{more extreme / skewed data}}$$

Example 1. *Suppose we have these data points:*

$$X = \{\ 1,\ 5,\ -1,\ 4,\ 6,\ 10,\ -4\ \}$$

Question: what are the min, max, mean, median*, standard deviation, and IQR?*

1. Reorder the data points:
$$\{\ 1,\ 5,\ -1,\ 4,\ 6,\ 10,\ -4\ \} \quad \longrightarrow \quad \{\ -4,\ -1,\ 1,\ 4,\ 5,\ 6,\ 10\ \}$$
2. Some metrics can be read off right away:
$$\{\ \underbrace{-4}_{\text{min}},\ -1,\ 1,\ \underbrace{4}_{\text{median}=Q_2},\ 5,\ 6,\ \underbrace{10}_{\text{max}}\ \}$$
3. mean
$$\text{mean} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}(-4 + -1 + \cdots + 10) = 3$$
4. standard deviation
$$\text{variance} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{6}\Big((-4-3)^2 + \cdots + (10-3)^2\Big) = \frac{1}{6}(132) = 22$$
$$\implies \text{standard deviation } = s = \sqrt{22} = 4.69$$
5. $IQR = Q_3 - Q_1$:
$$\{\ -4,\ \underbrace{-1,\ 1,}_{Q_1=0}\ 4\ \}; \qquad \{\ 4,\ \underbrace{5,\ 6,}_{Q_3=5.5}\ 14\ \} \qquad \implies IQR = Q_3 - Q_1 = 5.5 - 0 = 5.5$$

Notes. When working with discrete data (example above):

- Calculating Q_1 and Q_3 depends on conventions. For our class, if n is odd, we include the median $=$ Q_2 in calculating Q_1 and Q_3, like the example above.
- As such, different softwares might return different answers, but they should be very close.

Example 2. *Suppose we have the same data point except from the maximum point:*

$$Y = \{\ -4,\ -1,\ 1,\ 4,\ 5,\ 6,\ 500\ \}$$

Question: how are the $\min, \max, \text{mean}, \text{median}$*, standard deviation, and* IQR *changed?*

Some metrics are the same:

$$\min = -4, \quad Q_1 = 0, \quad \text{median} = Q_2 = 4, \quad Q_3 = 5.5, \quad IQR = 5.5$$

Some are different (hugely different):

$$\text{mean} = \bar{y} = 73, \qquad \text{standard deviation} = s_Y = 188.32$$

Notes:

- (median, IQR) remain unchanged while (mean, standard deviation) are changed by large margins.

$$\implies (\text{median}, IQR) \text{ are robust to outliers}$$