Republic of the Philippines
POLYTECHNIC UNIVERSITY OF THE PHILIPPINES

Elective 3
INTRODUCTION TO DATA SCIENCE
Home Activity

STATISTICAL INFERENCE

Name: Cueva, Larry Miguel, R.
Year and Section: BSCS 4-2

Homework / Home Activity Rubric

Predictor	Below Standard (2)	Approaching Standard (3)	At Standard (4)	Above Standard (5)
Completion	Student turned in assignment but mostly incomplete	Some of the assigned work is complete	Most of the assigned work is complete	All of the assigned work is complete
Accuracy	Little to none of the answers are correct	Some of the answers are correct	Most of the answers are correct	All of the answers are correct
Work Shown	Student did not show any work	Some steps for problem solving are missing	Most work is meticulously shown	All work is meticulously shown
Neatness	Homework is disorderly, with many smudges or tears	Homework is in a packet with several smudges or tears	Homework is in an orderly packet and is neat, with a few smudges or tears	Homework is in an orderly packet and is incredibly neat, with no smudges or tears

The rubric above will be used to evaluate your answer.

1. The following Temperatures in a City were recorded during a week in October: {78, 76, 71, 70, 68, 70, 71}

a. From the given temperatures above, find the **Mean** temperature.

```
> x_p_mean <- mean(x);
> x_p_mean
[1] 72
```

b. Find the **Median** temperature.

```
> x_median <- median(x);
> x_median
[1] 71
```

c. Find the **Mode**.

> x_mode <- names(table(x))[table(x) == max(table(x))];

> x_mode

[1] "70" "71"

d. Using a given dataset, create a histogram (using R) and calculate the standard deviation

> x_sd <- sd(x, na.rm=FALSE);

> x_sd

[1] 3.605551

e. Find the **Range**.

> x_range <- max(x)-min(x);

> x_range

[1] 10

f. Explain what does the **range** tell you?

Because the value is 10 this means that

g. What is the best **measure of center** and why?

Since the data is slightly skewed because of the smaller values having more occurrences than that of the larger more less occurring values on the right of the histogram, the best measure for central tendency is that of the median.

2. Consider influenza epidemics for two parent heterosexual families. Suppose that the probability is 17% that at least one of the parents has contracted the disease. The probability that the father has contracted influenza is 12% while the probability that both the mother and father have contracted the disease is 6%. What is the probability that the mother has contracted influenza?

$$P(Mother|Father) = P(Mother) + P(Father) - P(Mother\ and\ Father)$$

Where $P(Mother|Father)$ is the probability that either the mother or father contracts influenza, $P(Mother)$ the probability that the mother has contracted influenza, $P(Father)$ the probability that the mother has contracted influenza, and $P(Mother\ and\ Father)$ the probability that both the mother and father has contracted influenza. When values are plugged in the equation results in:

$$0.17 = P(Mother) + 0.12 - 0.06$$

$$-P(Mother) = -0.17 + 0.12 - 0.06$$

$$-P(Mother) = -0.11$$

$$P(Mother) = 0.11$$

Therefore probability that the mother has contracted influenza is 0.11

3. What is the difference between descriptive and inferential statistics? Please provide an example experimentation on your own, any field that is ***interesting to you*** and use R Programming Language. It

could be stated that the difference btween descriptive and inferential statistics is that descriptive statistics in of the word itself summarizes the characteristics or description of the data such as a data points tendency to be at the center, the maximum and minimum values of the data, the average of the data, the frequently occurring values in the data, these are all things that describe the data. For inferential statistics however it allows us to perhaps not much to describe the data but to hypothesize if such data we have is generalizable or can be concluded as something that can describe the behaviors, or characteristics of a population, as the name itself suggests, inferential means to infer, and to infer from the data given or to make a deduction about something based on data is the very definition of inferential statistics.

4. A bag contains 4 red balls and 6 green balls. If a ball is randomly selected from the bag, what is the probability that it is red?

$$P(red) = \frac{no.of\ red\ balls}{no.of\ all\ balls} = \frac{4}{10} = 0.4 = 40\%$$

5. The average score on a test is 85 with a standard deviation of 5. If a student receives a score of 92, what is their Z-score? Also, include your insights.

$$z = \frac{92 - 85}{5}$$

$$z = 1.4$$

6. A test is designed such that the mean score is 500 and the standard deviation is 100. If a student scores a 700, what is their Z-score? Then your insights?

$$z = \frac{700 - 500}{100}$$

$$z = 2$$

7. Do you think getting the measures of central tendency is enough to say that we have already the important facts about our dataset? Explain your answer. Yes, because from here we can compare even the most seemingly asymmetrical data in terms of the largeness of their value. For instance, if we had a dataset of scores that had a range in the hundreds, and another dataset with scores in the range between 1 and 10, then using the z score formula we can then normalize the larger values so that it can be comparable to that of the scores with the smaller values

8. A set of data has the following five numbers: 2, 4, 5, 9, and 11. Explain the step-by-step procedure on finding the interquartile range (IQR) of the data we have.

a. first is to get the lower quartile which is calculated by $(N + 1) * 0.25$

b. second is to get the middle quartile which is calculated by $(N + 1) * 0.5$

c. third is to get the upper quartile which is calculated by $(N + 1) * 0.75$

d. and lastly is to calculate the interquartile range by getting the difference between the upper

quartile and the lower quartile $Q3 - Q1$

Prepared by:
Montaigne G. Molejon,
MSIT CS/IT Instructor – PUP
CCIS