DA 6823

Kilger

Module 1: Part #2 (35 points)

**The Power of Statistics + the Levels of Measurement + the Different Classes of Variables and Determining Appropriate Statistical Technique + Basic Descriptive Measures**

1. The first couple of questions deal with the concepts of population and sample.
   1. What is a population? (3 points)

A population is a [set](https://en.wikipedia.org/wiki/Set_(mathematics)) of items that are grouped together by a common feature, which are of interest for some research or [experiment](https://en.wikipedia.org/wiki/Experiment). A population may refer to an entire group of people, objects, events, measurements, etc.

* 1. What is a sample? (3 points)

Sample is a set of data from a larger set called population, selected to represent the population.

Random sample is preferred to eliminate bias by giving all individuals/data points an equal chance to be chosen.

* 1. What is the objective of inferential statistics in terms of sample and population? (4 points)

The objective of inferential statistics is to draw conclusions about or generalize the entire population based on the sample, which is a subset of the same population. Random sample is preferred, to eliminate bias by giving all individuals/data points an equal chance to be chosen, so as to represent the population better.

1. Name two common measures of central tendency (4 points)

Mean and Median

1. Describe one situation where one measure of central tendency might be better than another measure (2 points)

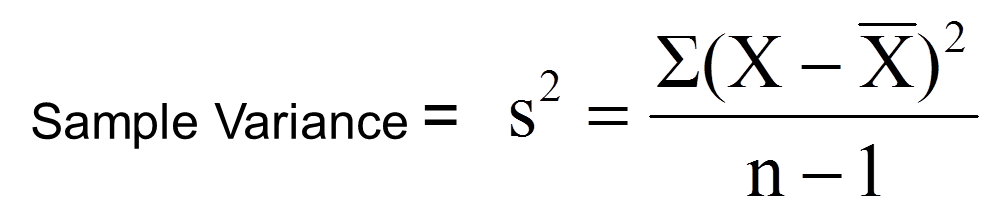
Median could be better than mean when data is skewed because median is robust to outliers.

For example, the mean salary for these ten staff is $30.7k But most of them have salaries in 12K to 18K range. This high mean is due to the 9th and 10th staff members(outliers). In such situations median(15.5) would be a better measure of central tendency

Staff 1 2 3 4 5 6 7 8 9 10

Salary 15k 18k 16k 14k 15k 15k 12k 17k 90k 95k

1. Variance and standard deviation are two of the most commonly used measures of variability. Take a look at the formula for variance below:



* 1. Looking at the variance formula, if the data points (X) are closely packed around the sample mean, what happens to the sample variance? (3 points)

If the distance between data points and their mean is small, the summation would be small and sample variance would be low compared to when the points are far apart from the mean.

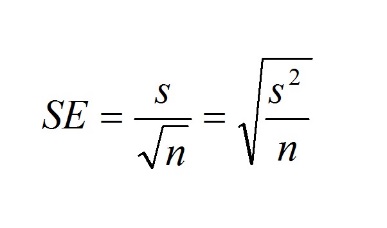
* 1. The sample standard deviation is closely related to the sample variance. How? (2 points)

Sample Standard deviation is square root of sample variance

1. Many sample statistics you encounter have standard errors associated with them. Imagine that you are studying the heights of the undergraduate student body of UTSA. The total number of students is 30,000 (e.g. the population) and you randomly pull 10 samples of 100 students each from that population.
   1. Will the sample means from each of the 10 samples be the same? ( 2 points)

No

* 1. The standard error of the mean represents the variation in sample means that you find in different samples. The formula below is the standard error of the mean.



1. What happens to the standard error of the mean if there is a lot of variation in the data? (3 points)

s is the standard deviation which is square root of variance. Both are directly proportional. So as variance increases s increases and so does the standard error.

ii. What happens to the standard error of the mean as sample size increases? (3 points)

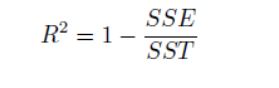
Since n is in the denominator as sample size increases standard error of the mean decreases

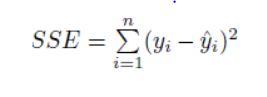
1. The sum of squares is a statistical concept that measures variation in data that you will find in many different statistical techniques. Here is the formula for sum of squares below:

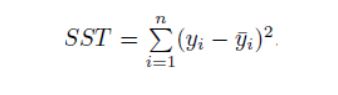


* 1. This formula should look familiar. What other statistical measures of variation are related to this? (3 points)

R-squared is related to sum of squares. It is the amount of variation in response explained by the model. The formula is:

where

 SSE is the sum of squares of regression model and SST is sum of squares of baseline model



* 1. As the data points (X) get spread out farther and farther from the sample mean, what happens to the sum of squares? (3 points)

As the data points spread farther the difference increases and its sum of squares increases too.