Introduction to Statistics

CHAPTER1

Statistics- is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, presenting and interpreting numerical information.

Descriptive statistics- utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present that information in a convenient form.

Inferential statistics- utilizes sample data to make estimates, decisions, predictions or other organizations about a larger set of data.

An experimental or observational unit is an object about which we collect

A population is a set of units that we are interested in studying.

A variable is a characteristic or property of an individual experimental unit in the population.

A sample is a subset of the units of population.

A statistical inference is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

A measure of reliability is a statement about the degree of uncertainty associated with a statistical inference.

4 elements of a descriptive statistical population 1) the population or sample of interest 2) one or more variables that are to be investigated 3) Tables, graphs, or numerical summary tools. 4) Identification of patterns in the data

5 elements of inferential statistical problem 1) The population of interest. 2) One or more variables that are to be investigated 3) The Sample of the population units. 4) the inference about the population based on information contained in the sample. 5) A measure of the reliability of the inference.

Quantitative data are measurements that are recorded on a naturally occurring numerical scale.

Qualitative data are measurements that cannot be measured on a natural numerical scale; they can only be classified into one of a group of categories.

A designed experiment is a data collection method where the researchers exert full control over the characteristics of the experimental units sampled. These experiments typically involve a group of experimental units that are assigned that treatment and an untreated group.

An observational study is a data collection method where the experimental units sampled are observed in their natural settings. No attempt is made to control the characteristics of the experimental units sampled.

A representative sample exhibits characteristics typical of those possessed by the target population.

A random sample of n experimental units is a sample selected from the population in such a way that every different sample of size n has an equal chance of selection.

Statistical thinking involves applying rational thought and the science of statistics to critically assess data and inferences. Fundamental to the thought process is that variation exists in populations of data.

Selection bias results when a subset of the experimental units in the population is excluded so that these units have no chance of being selected in the sample

Nonresponse bias results when the researchers conducting a survey or study are unable to obtain data on all experimental units selected for the sample.

Measurement error refers to inaccuracies in the values of the data recorded. In surveys, this kind of error may be due to ambiguous or leading questions and the inter-viewers effect on the respondent.

CHAPTER2

A class is one of the categories into which quantitative data can be classified.

The class frequency is the number of observations in the data set that fall into a particular class.

The class relative frequency is the class frequency divided by the total number of observations in the data set ; class relateive frequency = class frequency / n

The class percentage is the class relative frequency multiplied by 100; class percentage = class relative frequency x 100

Summary of graphical descriptive methods of qualitative data, Bar Graph: The categories classes of the qualitative variable are represented by bars, where the height of each bar is either the class frequency, class relative frequency, or class percentage. Pie Chart: The categories classes of the qualitative variable are represented by slices of pie circle. The size of each slice is proportional to the class relative frequency. Pareto Diagram: A bar graph with categories classes of the qualitative variable (ie the bars) arranged by height in descending order from left to right.

Summary of graphical descriptive methods of quantitative data, Dot Plot: the numerical value of each quantitative measurement in the data set is represented by a dot on a horizontal scale. When data values repeat, the dots are placed above one another vertically. Stem-and-Leaf Dislplay: the numerical value of the quantitative variable is partitioned in a “stem- and leaf”. The possible stems are listed in order in a column. The leaf for each quantitative measurement in the data set is placed in the corresponding stem row. Leaves for observation with the same stem value are listed in increasing order horizontally.

Hsitogram: the possible numerical values of the quantitative variable are partitioned into class intervals, each of which has the same width. These intervals form the scale of the horizontal axis. The frequency or relative frequency of observations in each class interval is determined. A vertical bar is placed over each class interval, with the height of the bar equal to either the class frequency or class relative frequency.

Numerical descriptive measures uses 2 methods: Central tendency is the tendency of data to cluster or center about certain numerical values. The variability of the set of measurements is the spread of the data.

The mean of a set of quantitative data is the sum of the measurements, divided by the number of measurements contained in the data set.

Symbols for the Sample Mean and the population mean. In this text, we adopt a general policy of using Greek letters to represent numerical descriptive measures of the population and Roman Letters to represent corresponding descriptive measures of the sample. Bar x is the symbol for Sample mean, mue greek character for population mean.

The median of the a quantitative data set is the middle number when the measurements are arranged in ascending order. When even number of measurements median is the mean of the middle 2 numbers.

A data set is said to be skewed if one tail of the distribution has more extreme observations than the other tail. If the data is skewed to the right then the median is less than the mean. If the data is skewed to the left, then the mean is less than the median. If the data is symmetric , then the mean equals the median.

The mode is the measurement that occurs most frequently in the data set. The measurement class containing the largest relative frequency is called the modal class.

Central tendency is partial and incomplete description of quantitative data, also needs measure of variability or spread.

The range of a quantitative data set is equal to the largest measurement minus the smallest measurement.

The sample variance for a sample of n measurements is equal to the sum of the squared deviations from the mean, divided by (n-1). The symbol s squared issued to represent sample variance.

The sample standard deviation, s, is defined as the positive square root of the sample variance, s squared.

S squared = sample variance; s = sampled standard deviation; sigma squared = population variance; sigma = population standard deviation.

Cheybyshevs rule applies to any data set , regardless of the shape of the frequency distribution data: It is possible that very few of the measurements will fall within one standard deviation of the mean; at least ¾ of the measurements will fall within 2 standard deviations of the mean; at least 8/9 of the measurements will fall within 3 standard deviations of the mean; generally, for any number k greater than 1, at least (1-1/k squared) of the measurements will fall within k standard deviations of the mean.

Empirical rule is the rule of thumb that applies to data set with frequency distributions that are mound shaped and symmetrical as follows: Approximately 68% of the measurements will fall within 1 standard deviation of the mean; approximately 95% of the measurements will fall within 2 standard deviations of the mean; approximately 99.7% of the measurements will fall within 3 standard deviations of the mean.

For an set of n measurements the pth percentile is a number such that p% of the measurements fall below that number and (100-p)% fall above it.

The lower quartile Ql is the 25th percentile of a data set. The middle quartile M is the median or 50th percentile. The upper quartile Qu is the 75th percentile.

The sample z score is x – sample mean / standard deviation. The population z-score is x – population mean (or meu)/ population standard deviation (or sigma). For a mound-shaped distribution of data: ~ 68% of the measurements will have z-score between -1 and 1;~ 95% of the measurements will have a z-score between -2 – 2; ~ 99.7% will have a z-score of -3 – 3.

An observation that is unusually large or small relative to the other values in a data set is called an outlier. Outliers typically are attributable to one of the following: The measurement is observed, recorded or entered into the computer incorrectly; the measurements comes from a different population; the measurement is correct, but represents a rare event.

Box plot is based on the quartiles of a data set. Specifically, a box plot is based on interquartile range, the distance between the lower and upper quartiles. Inner and outer fences or tails located +-1.5 IQR of hinges. Whiskers emanate from hinges. Whiskers extend to most observations at +-3 IQR and are deemed suspect outliers. Z-scores greater than 3 in abs value is an outlier. For slightly skewed data sets observations with z-score greater than 2 abs value are may be outliers.

The interquartile range or IQR is the distance between lower and upper quartiles.

Bivariate relationship describe relationship between 2 quantitative variables. Scatterplots, a 2 dimensional plot, is used by plotting one variables values in x axis and other in y axis. When increase in one variable is associated with an increase in the other they are said to be positively correlated. If one variable associated with negative they are negatively correlated.

Chapter 3

Probability is the opposite of statistics, while statistics uses a sample to infer the population , probability takes the population to infer the sample.

Mean/SD/IQR- StatCrunch Stats-Sumstats-Column or Applet(mean/SD/IQR)

z-score of 0 it is near the mean the larger positive measurement larger than most, large negative, smaller than most. > 2 possible outlier, >3 Outliers. Stats-Z-Stats-One Sample with summary. Look up z stat in Table III. Subtract from .5 for central theorem. Page 225-226

confidence level with Normal Z statistics- Stat- z-stat-One Sample with summary enter confidence level. Pg 245.

Histogram, Stat Crunch Graph – Historgram

Binomial – stats –calc- binomial Need x <>= p and n.

Standard Normal Table- falls between use the Table III score and multiply by 2. > use score and minus from .5 or .5 – score. < is score + .5 Use stat crunch stat – calc – Normal

T-Stats – Stat-T-Stat-One Sample With Summary.

Population Proportion- Stat- Proportion Stat one sample with Summary. PG 265

Regression – Stat – regression - simple linear.--- Confident interval If needed. Check for Coefficient correlation, r squared

Spearmans Correlation- Stat- Nonparametric test for correlation- spearmans correlations