STATISTICAL SOFTWARES THEORY NOTES

Definitions

Statistics
Collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.

Variable
Characteristic or attribute that can assume different values

Random Variable
A variable whose values are determined by chance.

Population
All subjects possessing a common characteristic that is being studied.

Sample
A subgroup or subset of the population.

Parameter
Characteristic or measure obtained from a population.

Statistic (not to be confused with Statistics)
Characteristic or measure obtained from a sample.

Descriptive Statistics
Collection, organization, summarization, and presentation of data.

Inferential Statistics
Generalizing from samples to populations using probabilities.Performing hypothesis testing, determining relationships between variables, and making predictions.

Qualitative Variables
Variables which assume non-numerical values.

Quantitative Variables
Variables which assume numerical values.

Discrete Variables
Variables which assume a finite or countable number of possible values.Usually obtained by counting.

Continuous Variables
Variables which assume an infinite number of possible values.Usually obtained by measurement.

Nominal Level
Level of measurement which classifies data into mutually exclusive, all inclusive categories in which no order or ranking can be imposed on the data.

Ordinal Level
Level of measurement which classifies data into categories that can be ranked. Differences between the ranks do not exist.

Interval Level
Level of measurement which classifies data that can be ranked and differences are meaningful. However, there is no meaningful zero, so ratios are meaningless.

Ratio Level

Level of measurement which classifies data that can be ranked, differences are meaningful, and there is a true zero. True ratios exist between the different units of measure.

Random Sampling

Sampling in which the data is collected using chance methods or random numbers.

Systematic Sampling

Sampling in which data is obtained by selecting every *k*th object.

Convenience Sampling

Sampling in which data is which is readily available is used.

Stratified Sampling

Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques.

Cluster Sampling

Sampling in which the population is divided into groups (usually geographically). Some of these groups are randomly selected, and then all of the elements in those groups are selected.

Population vs Sample

The population includes all objects of interest whereas the sample is only a portion of the population. Parameters are associated with populations and statistics with samples. Parameters are usually denoted using Greek letters (mu, sigma) while statistics are usually denoted using Roman letters (x, s).

There are several reasons why we don't work with populations. They are usually large, and it is often impossible to get data for every object we're studying. Sampling does not usually occur without cost, and the more items surveyed, the larger the cost.

We compute statistics, and use them to estimate parameters. The computation is the first part of the statistics course (Descriptive Statistics) and the estimation is the second part (Inferential Statistics)

Discrete vs Continuous

Discrete variables are usually obtained by counting. There are a finite or countable number of choices available with discrete data. You can't have 2.63 people in the room.

Continuous variables are usually obtained by measuring. Length, weight, and time are all examples of continous variables. Since continuous variables are real numbers, we usually round them. This implies a boundary depending on the number of decimal places. For example: 64 is really anything $63.5 <= x < 64.5$. Likewise, if there are two decimal places, then 64.03 is really anything $63.025 <= x < 63.035$. Boundaries always have one more decimal place than the data and end in a 5.

Levels of Measurement

There are four levels of measurement: Nominal, Ordinal, Interval, and Ratio. These go from lowest level to highest level. Data is classified according to the highest level which it fits. Each additional level adds something the previous level didn't have.

- Nominal is the lowest level. Only names are meaningful here.
- Ordinal adds an order to the names.
- Interval adds meaningful differences
- Ratio adds a zero so that ratios are meaningful.

Types of Sampling

There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified.

- Random sampling is analogous to putting everyone's name into a hat and drawing out several names. Each element in the population has an equal chance of occuring. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained. Computer generated lists are often used with random sampling.
- Systematic sampling is easier to do than random sampling. In systematic sampling, the list of elements is "counted off". That is, every *k*th element is taken. This is similar to lining everyone up and numbering off "1,2,3,4; 1,2,3,4; etc". When done numbering, all people numbered 4 would be used.
- Convenience sampling is very easy to do, but it's probably the worst technique to use. In convenience sampling, readily available data is used. That is, the first people the surveyor runs into.
- Cluster sampling is accomplished by dividing the population into groups -- usually geographically. These groups are called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.
- Stratified sampling also divides the population into groups called strata. However, this time it is by some characteristic, not geographically. For instance, the population might be separated into males and females. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

Statistic
: Characteristic or measure obtained from a sample

Parameter
: Characteristic or measure obtained from a population

Mean
: Sum of all the values divided by the number of values. This can either be a population mean (denoted by mu) or a sample mean (denoted by x bar)

Median
: The midpoint of the data after being ranked (sorted in ascending order). There are as many numbers below the median as above the median.

Mode
: The most frequent number

Skewed Distribution

The majority of the values lie together on one side with a very few values (the tail) to the other side. In a positively skewed distribution, the tail is to the right and the mean is larger than the median. In a negatively skewed distribution, the tail is to the left and the mean is smaller than the median.

Symmetric Distribution

The data values are evenly distributed on both sides of the mean. In a symmetric distribution, the mean is the median.

Weighted Mean

The mean when each value is multiplied by its weight and summed. This sum is divided by the total of the weights.

Midrange

The mean of the highest and lowest values. (Max + Min) / 2

Range

The difference between the highest and lowest values. Max - Min

Population Variance

The average of the squares of the distances from the population mean. It is the sum of the squares of the deviations from the mean divided by the population size. The units on the variance are the units of the population squared.

Sample Variance

Unbiased estimator of a population variance. Instead of dividing by the population size, the sum of the squares of the deviations from the sample mean is divided by one less than the sample size. The units on the variance are the units of the population squared.

Standard Deviation

The square root of the variance. The population standard deviation is the square root of the population variance and the sample standard deviation is the square root of the sample variance. The sample standard deviation is not the unbiased estimator for the population standard deviation. The units on the standard deviation is the same as the units of the population/sample.

Coefficient of Variation

Standard deviation divided by the mean, expressed as a percentage. We won't work with the Coefficient of Variation in this course.

Percentile

The percent of the population which lies below that value. The data must be ranked to find percentiles.

Quartile

Either the 25th, 50th, or 75th percentiles. The 50th percentile is also called the median.

Decile

Either the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, or 90th percentiles.
The difference between the 3rd and 1st Quartiles.

Outlier

An extremely high or low value when compared to the rest of the values.

Mild Outliers

Values which lie between 1.5 and 3.0 times the InterQuartile Range below the 1st Quartile or above the 3rd Quartile. Note, some texts use hinges instead of Quartiles.

Extreme Outliers

Values which lie more than 3.0 times the InterQuartile Range below the 1st Quartile or above the 3rd Quartile. Note, some texts use hinges instead of Quartiles

Measures of central tendency

The term "Average" is vague

Average could mean one of four things. The arithmetic mean, the median, midrange, or mode. For this reason, it is better to specify which average you're talking about.

Mean

This is what people usually intend when they say "average"

Population Mean:

$$\mu = \frac{\sum x}{N}$$

Sample Mean:

$$\bar{x} = \frac{\sum x}{n}$$

Frequency Distribution:

$$\bar{x} = \frac{\sum xf}{\sum f}$$

The mean of a frequency distribution is also the weighted mean.

Median

The data must be ranked (sorted in ascending order) first. The median is the number in the middle.

To find the depth of the median, there are several formulas that could be used, the one that we will use is:

```
Depth of median = 0.5 * (n + 1)
```

Raw Data

The median is the number in the "depth of the median" position. If the sample size is even, the depth of the median will be a decimal -- you need to find the midpoint between the numbers on either side of the depth of the median.

Ungrouped Frequency Distribution

Find the cumulative frequencies for the data. The first value with a cumulative frequency greater than depth of the median is the median. If the depth of the median is exactly 0.5 more than the cumulative frequency of the previous class, then the median is the midpoint between the two classes.

Grouped Frequency Distribution

This is the tough one.

Since the data is grouped, you have lost all original information. Some textbooks have you simply take the midpoint of the class. This is an over-simplification which isn't the true value (but much easier to do). The correct process is to interpolate.

Find out what proportion of the distance into the median class the median by dividing the sample size by 2, subtracting the cumulative frequency of the previous class, and then dividing all that bay the frequency of the median class.

Multiply this proportion by the class width and add it to the lower boundary of the median class.

$$MD = \tilde{x} = \left(\frac{n/2 - cf}{f} \right) w + L_m$$

Mode

The mode is the most frequent data value. There may be no mode if no one value appears more than any other. There may also be two modes (bimodal), three modes (trimodal), or more than three modes (multi-modal).

For grouped frequency distributions, the modal class is the class with the largest frequency.

Midrange

The midrange is simply the midpoint between the highest and lowest values.

Measures of Variation

Range

The range is the simplest measure of variation to find. It is simply the highest value minus the lowest value.

```
RANGE = MAXIMUM - MINIMUM
```

Since the range only uses the largest and smallest values, it is greatly affected by extreme values, that is - it is not resistant to change.

Variance

"Average Deviation"

The range only involves the smallest and largest numbers, and it would be desirable to have a statistic which involved all of the data values.

The first attempt one might make at this is something they might call the average deviation from the mean and define it as:

$$Ave.\ Dev\ = \frac{\sum\ (x\ -\ \mu)}{N}$$

The problem is that this summation is always zero. So, the average deviation will always be zero. That is why the average deviation is never used.

Population Variance

So, to keep it from being zero, the deviation from the mean is squared and called the "squared deviation from the mean". This "average squared deviation from the mean" is called the variance.

$$Population\ Variance = \sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Unbiased Estimate of the Population Variance

One would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

$$Sample\ Variance = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Standard Deviation

There is a problem with variances. Recall that the deviations were squared. That means that the units were also squared. To get the units back the same as the original data values, the square root must be taken.

$$Population\ Standard\ Deviation = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

$$Sample\ Standard\ Deviation = s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

The sample standard deviation is not the unbiased estimator for the population standard deviation.

The calculator does not have a variance key on it. It does have a standard deviation key. You will have to square the standard deviation to find the variance

Measures of Position

Standard Scores (z-scores)

The standard score is obtained by subtracting the mean and dividing the difference by the standard deviation. The symbol is z, which is why it's also called a z-score.

$$z = \frac{x - \mu}{\sigma} \quad or \quad z = \frac{x - \bar{x}}{s}$$

The mean of the standard scores is zero and the standard deviation is 1. This is the nice feature of the standard score -- no matter what the original scale was, when the data is converted to its standard score, the mean is zero and the standard deviation is 1.

Percentiles, Deciles, Quartiles

Percentiles (100 regions)

The kth percentile is the number which has k% of the values below it. The data must be ranked.

1. Rank the data
2. Find k% (k /100) of the sample size, n.
3. If this is an integer, add 0.5. If it isn't an integer round up.
4. Find the number in this position. If your depth ends in 0.5, then take the midpoint between the two numbers.

It is sometimes easier to count from the high end rather than counting from the low end. For example, the 80^{th} percentile is the number which has 80% below it and 20% above it. Rather than counting 80% from the bottom, count 20% from the top.

Note: The 50^{th} percentile is the median.

If you wish to find the percentile for a number (rather than locating the kth percentile), then

1. Take the number of values below the number
2. Add 0.5
3. Divide by the total number of values
4. Convert it to a percent

Deciles (10 regions)

The percentiles divide the data into 100 equal regions. The deciles divide the data into 10 equal regions. The instructions are the same for finding a percentile, except instead of dividing by 100 in step 2, divide by 10.

Quartiles (4 regions)

The quartiles divide the data into 4 equal regions. Instead of dividing by 100 in step 2, divide by 4.

Note: The 2^{nd} quartile is the same as the median. The 1^{st} quartile is the 25^{th}percentile, the 3^{rd} quartile is the 75^{th} percentile.

The quartiles are commonly used (much more so than the percentiles or deciles). The TI-82 calculator will find the quartiles for you. Some textbooks include the quartiles in the five number summary.

Hinges

The lower hinge is the median of the lower half of the data up to and including the median. The upper hinge is the median of the upper half of the data up to and including the median.

The hinges are the same as the quartiles unless the remainder when dividing the sample size by four is three (like 39 / 4 = 9 R *3*).

The statement about the lower half or upper half including the median tends to be confusing to some students. If the median is split between two values (which happens whenever the sample size is even), the median isn't included in either since the median isn't actually part of the data.

Example 1: sample size of 20

The median will be in position 10.5. The lower half is positions 1 - 10 and the upper half is positions 11 - 20. The lower hinge is the median of the lower half and would be in position 5.5. The upper hinge is the median of the upper half and would be in position 5.5 starting with original position 11 as position 1 -- this is the original position 15.5.

Example 2: sample size of 21

The median is in position 11. The lower half is positions 1 - 11 and the upper half is positions 11 - 21. The lower hinge is the median of the lower half and would be in position 6. The upper hinge is the median of the upper half and would be in position 6 when starting at position 11 -- this is original position 16.

Five Number Summary

The five number summary consists of the minimum value, lower hinge, median, upper hinge, and maximum value. Some textbooks use the quartiles instead of the hinges.

Box and Whiskers Plot

A graphical representation of the five number summary. A box is drawn between the lower and upper hinges with a line at the median. Whiskers (a single line, not a box) extend from the hinges to lines at the minimum and maximum values.

Interquartile Range (IQR)

The interquartile range is the difference between the third and first quartiles. That's it: `Q3 - Q1`

Outliers

Outliers are extreme values. There are mild outliers and extreme outliers. The Bluman text does not distinguish between mild outliers and extreme outliers and just treats either as an outlier.

Extreme Outliers

Extreme outliers are any data values which lie more than 3.0 times the interquartile range below the first quartile or above the third quartile. x is an extreme outlier if ...

```
x < Q1 - 3 * IQR
```

or

```
x > Q3 + 3 * IQR
```

Mild Outliers

Mild outliers are any data values which lie between 1.5 times and 3.0 times the interquartile range below the first quartile or above the third quartile. x is a mild outlier if ...

```
Q1 - 3 * IQR <= x < Q1 - 1.5 * IQR
```

or

```
Q1 + 1.5 * IQR < x <= Q3 + 3 * IQR
```

Correlation & Regression

Definitions

Coefficient of Determination

The percent of the variation that can be explained by the regression equation

Correlation

A method used to determine if a relationship between variables exists

Correlation Coefficient

A statistic or parameter which measures the strength and direction of a relationship between two variables

Dependent Variable

A variable in correlation or regression that can not be controlled, that is, it depends on the independent variable.

Independent Variable

A variable in correlation or regression which can be controlled, that is, it is independent of the other variable.

Pearson Product Moment Correlation Coefficient

A measure of the strength and direction of the linear relationship between two variables

Regression

A method used to describe the relationship between two variables.

Regression Line

The best fit line.

Scatter Plot

An plot of the data values on a coordinate system. The independent variable is graphed along the x-axis and the dependent variable along the y-axis

Standard Error of the Estimate

The standard deviation of the observed values about the predicted values