Statistics 201
Summary of Tools and Techniques

This document summarizes the many tools and techniques that you will be exposed to in STAT 201. The details of how to do these procedures is intentionally missing from this document. The purpose of this document is to prepare you to select the correct tool or technique for a specific situation. Knowing how to use each tool is important, but equally important is knowing the correct tool to use for a particular situation.

There will be several multiple choice questions on exam #3 that present you with a scenario, and offer 4-5 tools/techniques to choose from, and you must choose the correct one for that scenario. Refer to this document often throughout the semester, and make sure you understand the different types of problems that these tools and techniques are designed to address as we progress through the semester. It will make preparing for these questions on exam #3 much easier!

THE PICTURE TOOLS

Pie Charts, Bar Charts
When to use: to graphically display categorical data
Example: What search engine do internet users most frequently use?

Histograms
When to use: to graphically display quantitative variables, specifically to look at the shape, center, and spread of a distribution of data.
Example: What is the distribution of employee salaries at a public university in the SEC?

Stem and leaf plots
When to use: similar to a histogram, but allows user to recognize individual data points. Stem and leaf plots work best with small data sets.
Example: What is the distribution of the amount of money spent on school supplies for 20 families living in Knox County?

Pivot Table and Pivot Chart
When to use: to graphically summarize a combination of one or two categorical variables and one quantitative variable.
Example: What are the total sales for my 9 sales representatives, broken down by quarter.

Scatter Plots
When to use: to graphically display TWO quantitative variables, specifically to look for a possible linear relationship between the X-variable (independent/predictor) and Y-variable (dependent/predicted)
Example: Is there a linear relationship between credit scores and mortgage rates?

Mosaic Plots
When to use: to graphically display TWO categorical variables, specifically to look for a possible association (dependency).
Example: Is there an association between geographical location in the USA (northeast, southeast, mid-west, mid-Atlantic, etc) and political party alliance?

Box Plots
When to use: similar to histograms and stem-and-leaf displays, to graphically display a quantitative variable, display the median, the lower quartile and the upper quartile, as well as possible outliers. Side-by-side box plots are useful when comparing these characteristics of a quantitative variable at different "levels" of a categorical variable.
Example: Is there a difference (or similarity) between salaries of male associate professors when compared to salaries of female associate professors at the same college?

Decision Trees
When to use: To examine the relationship a y variable has when there are lots of x variables involved. A decision tree will find a set of best x variables to predict some y variable.
Example: What variables influence the price of a car? The data set contains 100 x variables collected from a dealership on their sales last month.

THE CALCULATION TOOLS

Frequency & Percent

When to use: to present a count, or percentage of a categorical variable.

Example: What percent (or count, or frequency) of first-born children attend college?

Mean, Median

When to use: calculate (or identify) these measures to describe the center of a distribution of a quantitative variable.

Mean: the average of all observations in a data set

Median: the observation in an ordered data set that divides the data into halves

Example: On average, how much money does a college freshman spend on text books?

Range, Interquartile Range, Standard Deviation

When to use: calculate (or identify) these measures to describe the spread (or variation) of a distribution of a quantitative variable.

Range: maximum value – minimum value

Interquartile range: the range of the middle 50% of the data

Standard deviation: the average deviation of all data points from the center of the distribution

Example: Describe the variation in the stock price of Amazon over the last 180 days.

r, (Pearson's correlation coefficient)

When to use: to assess the strength of a linear association between two quantitative variables.

Example: How strong is the linear association between high school GPA and college GPA?

R^2 (Coefficient of Determination)

When to use: to assess the amount of variation in the dependent (predicted) variable (y) that is associated with variation in the independent variable (x).

Example: What amount of the variation in movie budgets is associated with the variation in run time of movies?

slope, y-intercept (simple linear regression)

When to use: when you have two quantitative variables, and an equation of the form $\hat{y} = b_0 + b_1x$ models a linear relationship between these variables. This model can be used to make predictions for y at different values of x. The slope describes how the average value of y changes for different values of x. The intercept is an estimate of the average value of the y-variable when x equals zero (assuming x equals zero makes sense).

Example: For every additional gram of sugar in a serving of cereal, how many additional calories would we expect there to be, on average? How many calories do we expect, on average, in a serving of cereal that has 20 grams of sugar?

Calculating a sample proportion ($\hat{p}$)

When to use: When you have a sample of data, and you want to know the proportion of that sample that has the specific opinion (or trait, or outcome, etc.)

Example: What proportion of voters in my sample approve of the Affordable Healthcare Act? If you ask 100 people "Do you approve of the Affordable Healthcare Act?" and only 20 said yes; your sample proportion would be .20.

THE INFERENCE TOOLS

Constructing a Confidence Interval for a population proportion (p) from a sample proportion ($\hat{p}$)

When to use: You have calculated a sample proportion, and you want to give some boundaries on your estimation of the true population proportion.

Example: If your random sample of 100 people from Tennessee voters had 20 people answer yes to their approval of the Affordable Healthcare Act, the next step would be to construct a confidence interval for this estimate. A common confidence interval is a 95% confidence interval, which would be ($0.1216 \leq p \leq .2784$). We can be 95% confident that the true proportion of Tennessee voters that approve of the Affordable Healthcare Act is between 0.122 and 0.278.

Comparing a sample proportion ($\hat{p}$) to a population proportion (p) by hypothesis testing

When to use: You have an idea of what your population proportion is or should be, generally based on past research or historical data, and you want to investigate if this population proportion is valid, or has there been a change.

Example: When the Affordable Healthcare Act was introduced in 2012, it was claimed that 55% of registered voters approved of it. A sample of 100 registered voters shows that only 20% of this sample approves of the Affordable Healthcare Act. Is this unusual, if the original claim is true?

Constructing a Confidence Interval for a population mean (μ) from a sample mean ($\bar{y}$)

When to use: You have calculated a sample mean, and you want to give some boundaries on your estimation of the true population mean.

Example: If your random sample of 100 recent UT graduates had an average of $10,000 in student loan debt, and a $1000 standard deviation, a common confidence interval is a 95% confidence interval, which would be ($\$9{,}802 \leq \mu \leq \$10{,}198$). We can be 95% confident that the true average amount of student loans of UT graduates is between $9802 and $10,198.

Comparing a sample mean ($\bar{y}$) to a population mean (μ) by hypothesis testing

When to use: You have an idea of what your population average is or should be, generally based on past research or historical data, and you want to investigate if this population mean is valid, or has there been a change.

Example: Kiplinger magazine reported that the average student loan debt for a graduate of the University of Tennessee is $9,600. Our sample of 100 recent UT grads showed an average of $10,000 with a standard deviation of $1000. Is this sample result unusual, if the Kiplinger magazine claim is true?

Constructing a Confidence Interval for the true difference between two population means (μ_1 - μ_2) from a sample difference between two means ($\bar{y}_1 - \bar{y}_2$)

When to use: You have two samples of quantitative data that are independent of each other. You have calculated a difference between two sample means, and you want to give some boundaries on your estimation of the true difference between the two means.

Example: We randomly sample 100 recent UT graduates and find an average of $10,000 in student loan debt with a $1000 standard deviation. We also randomly sample of 100 recent Florida graduates and find an average of $10,500 in student loan debt with a $2000 standard deviation. A common confidence interval is a 95% confidence interval, which would be ($440 ≤ μ_1 - μ_2 ≤ $560). We can be 95% confident that the true difference between the average amount of student loads between UT graduates and Florida graduates is between $440 and $560 more for Florida graduates.

Comparing a difference in two independent sample means ($\bar{y}_1 - \bar{y}_2$) to a hypothesized difference (usually 0) by hypothesis testing.

When to use: You have two samples of quantitative data that are independent of each other. You wish to test to see if the true different between the two population means is "statistically significant" (i.e., unlikely to be zero).

Example: Incoming freshman are asked to take two different tests to assess their math abilities. The tests claim to test the same knowledge and should have the same average score. Twenty five randomly selected students are asked to take the first test and another twenty five randomly selected students are asked to take the second test. The mean and standard deviation for the first test are 88 and 4 respectively, while the mean and standard deviation for the second test are 90 and 5 respectively. Do these results lead to evidence that there is a statistically significant difference between the population means of these two tests?

Chi-square test of independence

When to use: Data for two categorical variables are collected from a population of interest. The researcher wants to investigate if these two categorical variables are independent of each other.

Example: As far as UT students are concerned, is gender and whether you live on or off campus independent of each other?