# CS130/230 Lecture 13
# Hypothesis Testing

Tuesday, March 16, 2004

## *Problem 1*

The following table gives data on the lean body mass (kilograms) and the resting metabolic rate for 12 women and 7 men who are subjects in a study of obesity. The researchers suspect that lean body mass (that is, the subject's weight leaving out all fat) is an important influence on metabolic rate.

| Subject | Gender | Mass | Rate |
|---|---|---|---|
| 1 | m | 62.0 | 1792 |
| 2 | m | 62.9 | 1666 |
| 3 | f | 36.1 | 995 |
| 4 | f | 54.6 | 1425 |
| 5 | f | 48.5 | 1396 |
| 6 | f | 42.0 | 1418 |
| 7 | m | 47.4 | 1362 |
| 8 | f | 50.6 | 1502 |
| 9 | f | 42.0 | 1256 |
| 10 | m | 48.7 | 1614 |
| 11 | f | 40.3 | 1189 |
| 12 | f | 33.1 | 913 |
| 13 | m | 51.9 | 1460 |
| 14 | f | 42.4 | 1124 |
| 15 | f | 34.5 | 1052 |
| 16 | f | 51.1 | 1347 |
| 17 | f | 41.2 | 1204 |
| 18 | m | 51.9 | 1867 |
| 19 | m | 46.9 | 1439 |

(a) Enter this dataset in StatView and save it as: **obesity.svd** where you are saving all of your other files.

(b) Create a Cell Bar Chart containing both Mass and Rate divided by Gender. Title your graph "Mean Gender Study: Mass & Rate". Paste this into your document.

(c) Construct a Bivariate Regression Plot of the association between Mass and Rate for the entire group. Paste this into your document with a title of "Bivariate Regression Plot". Make sure you've properly selected the dependent and independent variable.

(d) What would you expect your Rate to be for a Mass of 45.25?

## Hypothesis Testing

Hypothesis testing is a decision making process for evaluating claims about a population.

The researcher must:

- Define the population under study
- State the hypothesis that is under investigation
- Give the significance level
- Select a sample from the population
- Collect the data
- Perform the statistical test
- Reach a conclusion

Examples include z-test, t-test, and correlation analysis to name a few.

My goal is to give you enough information to use StatView to perform some different hypothesis tests without getting into the nitty gritty detail. To use this tool properly, you must have a statistics background; otherwise, chances are an incorrect test can be applied to a given set of data to name one problem.

Hypothesis testing is the formal statistical technique of collecting data to answer questions through the use of a statistical model.

Each question asked needs to be posed as a null hypothesis; the null hypothesis is that there are no differences of the dependent variables of your model that can be explained by the independent variables.

## ONE-SAMPLE T-TEST:

This is the easiest of the statistical tests to understand. Specifically, this test compares a sample mean (computed from a set of observed values) to a hypothesized mean and determines the likelihood that the observed difference between the sample and hypothesized mean occurs by chance. The chance is reported as the p-value.

A p-value close to 1 means that it is very likely that the hypothesized and sample means are the same (assuming that they are the same), and a small p-value (for example 0.01) means it is unlikely (only a one in 100 chance) that such a difference would occur by chance (again, assuming that the two means are the same).

Thus, the lower the p-value the more certain that we can be that there is a statistically significant difference between the observed and hypothesized mean. Most disciplines use an alpha value of 0.05; that is, if the p-value is less than 0.05 then the difference is regarded as statistically significant.

- if $p < 0.05$, reject the null hypothesis
- if $p >= 0.05$, accept the null hypothesis

## *Problem 2*

Our first objective is to perform a one-sample t-test on data from blood lipid screening of medical students. Specifically, we will test whether the mean cholesterol level is significantly greater than 190, the point at which cholesterol levels may be unhealthy. You will test the null hypothesis that the mean value for cholesterol is 190.

**Null hypothesis:** The mean value for cholesterol is 190.

- Open Lipid Data.
- From the Analyze menu, select NEW VIEW.
- In the analysis browser, select ONE SAMPLE ANALYSIS and check CREATE ANALYSIS
- For a hypothesis mean, type 190 and click OK
- In the variable browser, select Cholesterol and click ADD

You will notice that the mean is slightly higher than the hypothesized value of 190. However, although the mean is in fact higher, this difference is well within the range of sampling variance. In particular, the p-value of 0.7373 indicates you would see a difference of this magnitude by chance more than 73% of the time. Thus the mean cholesterol level is not significantly greater than 190.

## PAIRED T-TEST:

The most common use of the paired t-test is the comparison of two measurements (typically one measurement occurs before a "treatment" and the other after the "treatment") from the same individual or group. This test is used primarily to determine if the "treatment" had a statistically significant effect. As in the case of the one-sample t-test, the primary statistic of concern is the p-value, and thankfully the p-value has the same interpretation here as it did in the case of the one-sample t-test.

## *Problem 3*

Again using the Lipid Data file we want to see if there is a statistically significant difference between Triglycerides and Trig-3.

**Null hypothesis:** There is no statistical difference between the two values tested.

- Open the Lipid Data file used in example one.
- From the Analyze menu, select NEW VIEW.
- In the analysis browser, select PAIRED COMPARISONS and check CREATE ANALYSIS.
- Click OK to accept the default parameters
- In the variable browser, select Triglycerides and Trig-3 yrs and click ADD.

**Question:** Should we accept the null hypothesis in this case?

## Z-TEST

This test is somewhat similar to the paired t-test, but this test is used to determine if there is a relationship between two unrelated measurements (as opposed to similar measurements in the paired t-test), and this is done by way of the correlation coefficient.

Typically, the hypothesized correlation coefficient is 0 (the default), and the statistics of concern are the correlation coefficient and the p-value. Recall that a correlation coefficient of either 1 or -1 is the ideal, and a correlation coefficient of 0 indicates that there is no correlation between the two variables. The p-value again has the same interpretation as it relates to the correlation coefficient.

### *Problem 4*

Null hypothesis: There is no difference in climate & terrain that can be explained by housing.

- Open the Western States Rated file from the Sample Data folder.
- From the Analyze menu, select NEW VIEW.
- In the analysis browser, select PAIRED COMPARISONS and check CREATE ANALYSIS.
- Uncheck Paired t-test and check Z-test under Correlation. Make sure that you leave the hypothesized correlation set to 0 as we are testing the hypothesis of no relationship between the variables.
- In the variable browser, select Climate& Terrain and click ADD.
- In the variable browser, select Housing and click ADD.

**Question:** Is there a correlation between Climate & Terrain and Housing? How do we know?

**Question:** Should we accept the null hypothesis?

## UNPAIRED T-TEST:

This test is very similar to the paired t-test, but instead of comparing two measurements within our entire population, we use only one measurement but break our population into two natural subgroups, testing whether there is a statistically significant difference between the means of these two subgroups. As in the case of the paired t-test, our primary statistic of concern is the p-value, and again it has the same interpretation.

## *Problem 5*

Null hypothesis: There is no difference between the price of houses near to and far from the Charles River.

- Open the Boston Housing Data file from the Sample Data folder.
- From the Analyze menu, select NEW VIEW.
- In the analysis browser, select UNPAIRED COMPARISONS and check CREATE ANALYSIS.
- Click OK to accept the default parameters.
- In the variable browser, select Median Value and click ADD.
- In the variable browser, select Charles and click ADD.

Question: Should we accept the null hypothesis?

## *Problem 6*

The dean of students wants to see whether there is a significant difference in age between resident students and commuting students. She selected a random sample of 10 students from each group.

The data follows:

| Resident Students | Commuter Students |
|---|---|
| 22 | 18 |
| 25 | 20 |
| 27 | 19 |
| 23 | 18 |
| 26 | 22 |
| 28 | 25 |
| 26 | 24 |
| 24 | 22 |
| 25 | 23 |
| 20 | 18 |

**Null hypothesis:** There is no statistical significant difference in the age between resident and commuting students in the above study.

**Question:** Should we accept the null hypothesis?