Hypothesis Testing

Fall 2014

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

Population on Samples

- Population: the entire collection of individuals about which information is sought
- Sample: subset of a population, containing the individuals that are actually observed

Population and Samples

Give at least three examples of a population

1.

2.

3.

For the population listed in 1., give an example of a sample from the population

Can you make up some hypothesis about the population in 1.

Hypothesis Tests

- Examples of hypothesis tests include t-test, Chi-Square, and correlation analysis to name a few
- To use this tool properly, you must understand the statistics
- Applying an incorrect test to a given set of data will give incorrect results

Hypothesis Testing

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

- "In statistics, a result is called **statistically significant** if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the **significance level**."

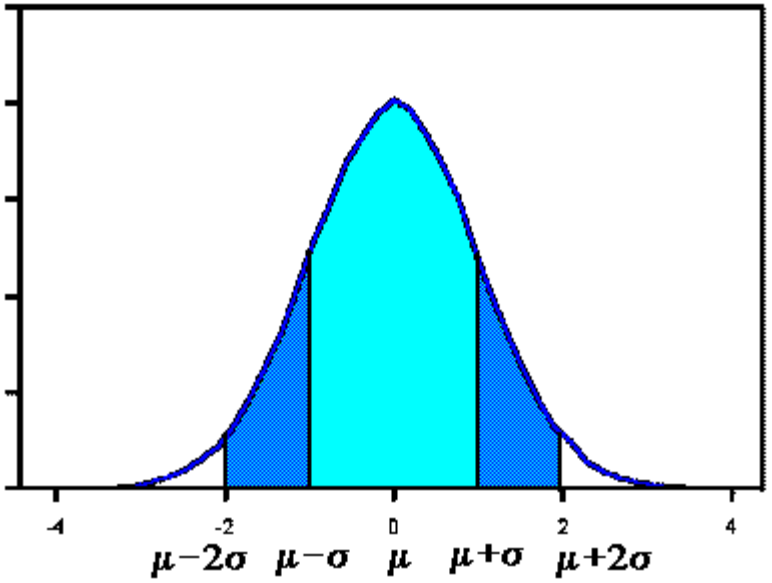
  http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

NULL Hypothesis

- The null hypothesis refers to a general or default position – denoted H_0
- Null hypothesis is assumed true until evidence indicate otherwise

The Normal Distribution

- The following Hypothesis Tests assume that the data is **normally distributed**.
- The standard normal curve in the picture has a mean of 0 and standard deviation of 1. A dataset with a normal distribution has about 68% of the observations within σ of the mean μ which in this case is (-1,1)



http://www.stat.yale.edu/Courses/1997-98/101/normal.htm

The Normal Distribution Continued

- About 95% of the observations will fall within 2 standard deviations of the mean (-2,2)
- About 99.7% of the observations will fall within 3 standard deviations of the mean
- Example: Consider 130 observations of body temperature with the results below. If the data is normal, what must be the case?

Variable	N	Mean	Median	StDev	Min	Max
BODY TEMP	130	98.249	98.300	0.733	96.300	100.800

Hypothesis Tests

- We will be using the following hypothesis tests in this course:
  - One sample t-test
  - Unpaired or independent samples t-test
  - Paired t-test
  - Correlation analysis

One-Sample T-Test

- This is the easiest of the statistical tests to understand

- Compare **observed** vs **hypothesized** mean
  - Observed: measured
  - Hypothesized: we choose this value to be meaningful
- T-Test determines the likelihood that the difference between the means occurs by chance
- The chance is reported as the p-value

p-value

- p-value: the probability that the difference occurs due to chance
  - A small p-value means that the difference is unlikely to be the result of chance
  - A large p-value means the difference is likely to be the result of chance
- What do we mean by random chance? Keep this question in mind and we will come back and give an answer.

Statistically Significant Difference

- The lower the p-value, the more certain that we can be that there is a **statistically significant** difference

- Most disciplines look for a p-value of 0.05 or less
  - if $p < 0.05$, reject the null hypothesis
  - if $p >= 0.05$, do not reject the null hypothesis

Problem 11.1

The file LipidData in the CS130 Public directory represents a blood lipid screening of medical students.

1. Grab this Excel file, open it up in SPSS and save the file as lipiddata.sav.
2. What is the mean Cholesterol value?
3. Is the cholesterol level significantly greater than 190? Can you tell by looking at the data? What do you think?

Problem 11.1 Continued

- 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 greater than 190 in a *statistically significant* way, the point at which cholesterol levels may be unhealthy.

What is the NULL hypothesis?

What is the alternative hypothesis?

Problem 11.1 Continued

1. Open Lipid Data.
2. From the Analyse menu, select Compare Means and then One Sample t-test .
3. Select your Test Variable which is Cholesterol .
4. Enter the Test Value which is 190.
5. In the variable browser, select Cholesterol and click ADD

Problem 11.1 Results

- The p-value is given in the box labeled Sig. (2-tailed) which stands for significance level

One-Sample Test

	Test Value = 190					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Cholesterol	.336	94	.737	1.23158	-6.0356	8.4988

Problem 11.1 Results

- The mean is slightly higher than 190; however, this difference is well within the range of sampling variance.
- A significance level of .737 indicates you would see a difference of this magnitude by chance more than 73% of the time
- Thus the 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" a treatment from the **same** individual or group.
- This test can determine if the treatment had a statistically significant effect.
- The p-value is the primary statistic of concern and the interpretation of the p-value is the same as for the one-sample t-test

Problem 11.2

- Using the LipidData

  1. What is the mean for Triglycerides?
  2. What is the mean for Trig-3yrs?
  3. Does it look like there is a statistically significant difference between Triglycerides and Trig-3yrs?

Problem 11.2 Continued

- Perform the paired t-test using the LipidData file
- State the Null Hypothesis and the alternative hypothesis
- From Analyze menu, select Compare Means and then Paired Samples t-test
- Should we accept the Null Hypothesis? Why?
- State your conclusion

Unpaired T-Test

- One measurement per individual
- Break our population into two natural subgroups
  - Male/Female; Smoker/Non-Smoker; Oak/Maple
  - Do the groups have a difference in measurement?
- Our primary statistic of concern is the p-value
  - How likely to occur by chance?

Problem 11.3

Question: Are the prices of houses near the Charles River more expensive than the prices of houses away from the Charles River.

The file BostonHousingData in the CS130 Public directory contains information about Boston houses.

1. Grab this Excel file, open it up in SPSS and save the file as bostonhousingdata.sav.

2. State the Null Hypothesis and the alternative hypothesis

3. Perform an unpaired t-test (Independent Samples T-Test in SPSS)

Problem 11.3

- What is the test variable? Why?
- What is the grouping variable? Why
- Next, Define Groups

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Median Value	Equal variances assumed	8.752	.003	-3.996	504	.000	-6.34616	1.58795	-9.46598	-3.22633
	Equal variances not assumed			-3.113	36.876	.004	-6.34616	2.03841	-10.47683	-2.21548

- The columns labeled "Levene's Test for Equality of Variances" tell us whether an assumption of the t-test has been met. The t-test assumes that the variability of each group is approximately equal.
-
- Look at the column labeled "Sig." under the heading "Levene's Test for Equality of Variances". In this example, the significance (p value) of Levene's test is .003.
  - If this value is less than or equal to 0.05, then you should use the bottom row of the output (the row labeled "Equal variances not assumed.")
  - If the value is greater than .05, then you should use the top row of the output (the row labeled "Equal variances assumed.")
  - In this example, .003 is lower than .05, so we will assume that the variances are not equal and we will use the bottom row of the output.

Problem 11.3

- Do you accept or reject the Null Hypothesis? Why?
- State your conclusion

Correlation Analysis

- Correlation Analysis addresses the following: Is there a statistically significant association between variable X and variable Y?
- Interpreting the Pearson Correlation Coefficient is not an exact science. We might use the following interpretation:
  - -1.0 to -0.7 strong negative association
  - -0.7 to -0.3 weak negative association
  - -0.3 to +0.3 little or no association
  - +0.3 to +0.7 weak positive association
  - +0.7 to +1.0 strong positive association

Correlation Analysis Visual

- Use Scattergrams (Scatterplots) to visually display data analyzed with this test.

- You can also produce a correlation matrix of the relationship of all variables in the matrix.
- Analyze | Correlate | Bivariate

Problem 11.4

- Create a correlation matrix of Cholesterol, Triglycerides, HDL, and LDL.

Identify the strongest positive correlation in the matrix.

Analyze | Correlate | Bivariate