Quiz 4 Review

171:161 Intro to Biostatistics

Deciding the Test to use…

- Categorical data:
  - Use a Pearson's Chi-Square or Fisher's Exact test
  - If it is a single sample use a binomial test
- Normally distributed continuous data:
  - Use a two sample t-test… Student's test if the variances are equal, Welch's if they aren't
- Continuous data that is skewed/not normal:
  - Try a transformation
  - Use a non-parametric approach… Wilcoxon Signed Rank if paired/matched groups, Wilcoxon Rank Sum if independent groups

Ex #1: A study comparing the miles per gallon of American Cars (sample 1) vs. Japanese Cars (sample 2) obtained the following results:

SAMPLE 1:

NUMBER OF OBSERVATIONS	= 249
MEAN	= 20.14458
STANDARD DEVIATION	= 6.41470

SAMPLE 2:

NUMBER OF OBSERVATIONS	= 79
MEAN	= 30.48101
STANDARD DEVIATION	= 6.10771

Part A) Assuming data are normal, what test should we use?

Ex #1: Part B)

The pooled SD here is 6.34260. Conduct a t-test comparing the mean MPG of Japanese vs. American cars.

Part B) ANSWER

$$T_{obs} = \frac{20.14 - 30.48}{6.34\sqrt{\frac{1}{249} + \frac{1}{79}}} = -12.62$$

Using (249+79 – 2) = 326 degrees of freedom,

the p-value is less than 0.05, there is a difference in MPG, American cars get fewer miles per gallon.

Ex #1 Part C)

Suppose the American Car groups has 2 observations getting 4, and 5 MPG. How might this effect the test results?

What statistical methods might you consider to remedy this?

Part C) ANSWER

The t-test procedure can be effected by impact outliers have on the mean and standard error.

Assuming the measurements are accurate you might consider a non-parametric (Wilcoxon Rank Sum) approach to remove the effect of these outliers.

You can also consider a log transformation to shrink the scale of the variables.

Note: A confidence interval done on log-transformed data will reflect the ratio group means after being exponentiated.

Two Sample Categorical Data

- We have looked at categorical data before using the Binomial test, which applied to single samples or paired samples.
- Two sample techniques are needed when:
  - We don’t have paired observations
  - We have observational data on an outcome and an exposure. (Why must this be treated as two sample?)

Contingency Tables

Example #2:

In 2000 Vermont State legislature approved a bill authorizing civil unions. Below is a summary of the legislators by gender and vote.

35 of the 44 female legislators supported the bill.

60 of the 101 male supported the bill.

Part A) Construct a 2x2 contingency table displaying the data.

	"Yes" Voter	"No" Voter	Total
Women	35	9	44
Men	60	41	101
Total	95	50	145

- Typically the outcome is shown in the table's columns and the exposure/treatment group is shown in the rows table's rows.
- However, having things flipped will not change the results of hypothesis testing!

	"Yes" Voter	"No" Voter	Total
Women	35	9	44
Men	60	41	101
Total	95	50	145

Part B) Suppose we are interested in whether there is an association between gender and support towards gay marriage. What is an appropriate Null Hypothesis?

Create a table showing the expected cell counts under the Null Hypothesis of no association.

ACTUAL COUNTS	"Yes" Voter	"No" Voter	Total
Women	35	9	44
Men	60	41	101
Total	95	50	145
EXPECTED COUNTS	**"Yes" Voter**	**"No" Voter**	**Total**
Women	28.83	15.17	44
Men	66.17	34.83	101
Total	95	50	145

Expected counts can be found by multiplying that cell's row total * the cell's column total, and dividing by the grand total

For example the Women, Yes cell has an expected count of:

$$\frac{44*95}{145} = 28.83$$

Part C) Conduct a test to determine if there is an association at the $\alpha = 0.05$ level.

Part C) ANSWER When doing a test by hand we will use a Pearson's Chi-Square Test. The test statistic is:

$$\chi^2 = \frac{\Sigma(obs - expect)^2}{expect}$$

$$\chi^2 = 5.50$$

The degrees of freedom are found by:

(R – 1)*(C – 1) = 1

The p-value here is 0.02

The conclusion is that there is an association, specifically that men were less supportive of the bill than women were.

Part D) Could we have used Fisher's Exact Test here? In what situations does Fisher's Test need to be used instead of the Pearson Chi-Square test?

- Yes Fisher's Exact Test can be used for any categorical table. It is an exact approach and the Chi-Square test is an approximate approach.
- Fisher's test does much better when cell counts are small. Generally expected cell counts of less than 5 can start to cause problems for the Chi-Square test.
- With large samples the two approaches will produce similar results.

Quantifying Association for 2 sample Categorical Data

- Differences in proportions can often be misleading in many public health situations.
- Because this Relative Risk and Odds Ratios are most often used to report the strength of an association.

$$RR = \frac{P(D|Exp)}{P(D|Unexp)}$$

$$OR = \frac{Odds(D\ among\ Exp)}{Odds(D\ among\ Unexp)} = \frac{a * d}{b * c}$$

	"Yes" Voter	"No" Voter	Total
Women	35	9	44
Men	60	41	101
Total	95	50	145

Ex 2: Part E) Find an interpret the odds ratio for the contingency table.

Ex 2: Part E) ANSWER

$$OR = \frac{a * d}{b * c} = \frac{35 * 41}{9 * 60} = 2.65$$

So the odds of a female representative supporting civil unions are 2.65 times the odds of a male representative supporting civil unions.

Odds vs. Relative Risk

- Relative Risk doesn't make sense for Case-Control retrospective studies.
- Relative Risk is not symmetric ("risk of dying without treatment" is not the same as "risk of surviving with treatment"
- Odds ratios are symmetric

Confidence Intervals for Odds Ratios

- CI's must be made on the Log scale since odds ratios are bounded below by 0

$$CI = \log(OR) \pm Z_{\alpha/2} * \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Exponentiate the upper and lower limits calculated above to obtain the CI for the Odds Ratio