

MSDS 6371 Exam 1 – Summer 2015

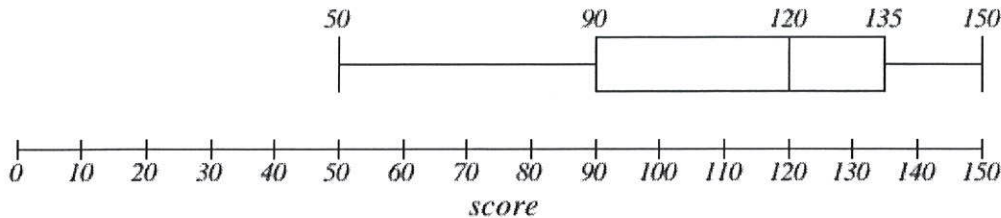
Answer the questions that follow to the best of your ability. On questions that require hand calculations, please show the formula used and the formula with the correct numbers in the correct places in order to get full credit for the problem. For multiple choice questions 1,2,4, you may simply highlight the correct answer or make it a different color. For short answer questions, answers should be in the 2-4 sentences range.

Use the boxplots below to answer questions 1 and 2.

Midterm 1



Midterm 2



1. (4 points) The boxplot above shows the grades of students in a statistics class on two midterms. Which midterm has a greater percentage of students with scores at or above 120?
  - a. Midterm 1
  - b. Midterm 2
  - c. Both Midterms are about equal
  - d. It is impossible to tell this level of detail from a boxplot
2. (4 points) Refer again to the boxplot above. Which of the following is correct?
  - a. The means of both midterms are larger than their medians
  - b. The means of both midterms are smaller than their medians
    - i. Assume 120 is median, left shift indicates mean is likely smaller than median
  - c. The means of both midterms are about the same as their medians
  - d. There is no way to tell the relationship between mean and median from a boxplot
3. (4 points) What does it mean to say that a result is statistically significant?
  - a. A result is statistically significant if it is caused by something other than random chance. This is explained with a p-value normally in a formal hypothesis test. This is different from practical significance and both should be considered during a formal hypothesis test.

\* p values: calculated assuming the null is true.

4. (4 points) An agricultural researcher plant 25 plots with a new variety of corn that is drought resistant and hence potentially more profitable. The average yield for these plots is 150 bushels per acre. Assume that the yield per acre for the new variety of corn follows a normal distribution with unknown mean  $\mu$  and that a 95% confidence interval for  $\mu$  is found to be 150  $\pm$  3.29. Which of the following is true?

- a. A test of the hypotheses  $H_0: \mu = 150$ ,  $H_a: \mu < 150$  would be rejected at the 0.05 level
- b. A test of the hypotheses  $H_0: \mu = 150$ ,  $H_a: \mu \neq 150$  would be rejected at the 0.05 level
- c. A test of the hypotheses  $H_0: \mu = 150$ ,  $H_a: \mu > 150$  would be rejected at the 0.05 level
- d. A test of the hypotheses  $H_0: \mu = 160$ ,  $H_a: \mu \neq 160$  would be rejected at the 0.05 level

5. (12 points) You have recently taken a job at a research facility, and your first duty is to calculate a sample size for a study. You type the following program into SAS

```
proc power;  
onesamplemeans  
mean = 3  
nullmean = 0  
ntotal = .  
stddev = 10  
power = .8; run;
```

- a. What is the value of the probability of Type II error?

a. 0.196

$$1 - 0.8 = 0.2$$

- b. What is the value of the probability of Type I error?

a. 0.05

- c. Suppose the standard deviation is decreased to 8. What will happen to the number of subjects, all else staying the same?

- a. The amount of subjects will decrease. There is less variation to account for so a smaller sample size can be acquired.

6. (4 points) Suppose a researcher writes in a journal article that "the obtained p was  $p = 0.032$ ; thus, there is only a 3.2% chance that the null hypothesis is correct." Is this a correct or incorrect statement? Explain your answer.

- a. The statement is incorrect. A p value is the probability of obtaining a test statistic under the null when the null is true, not the probability that the null hypothesis is correct. A p value should be interpreted as evidence against the null hypothesis if it is low (under your threshold) and practical.

after good!

\* null is either true or not true.

7. **Presentation counts! Keep your analysis to 2 pages (Single sided) including graphs, plots and charts. There should be 1 page max for each statistical test (2 tests total.) Include all statistical symbols such as  $\mu$  and  $\sigma$ . Finally, remember that a "Test" includes addressing the assumptions, doing the necessary statistical analysis, and writing a meaningful conclusion.**

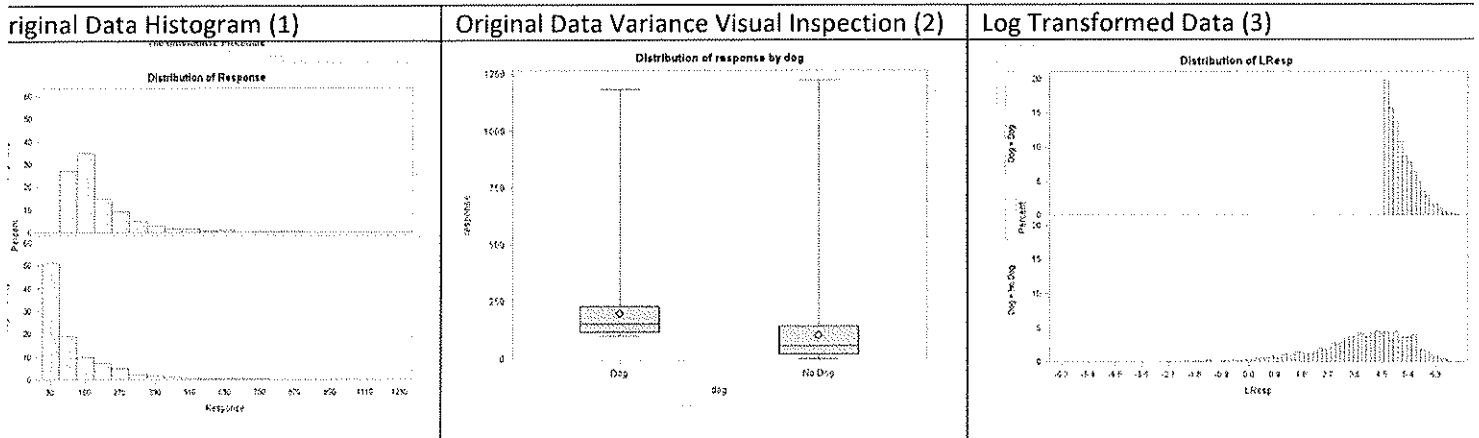
The data consist of 3974 happiness values recorded from a random sample of 3974 people from Missouri. The higher the score the happier the subject is reported to be. 1996 of the people had a dog (and were thus put in the 'dog' group) and 1978 did not have a dog (and were thus put into the "No Dog" Group.) The researcher would like to know if the happiness scores of the dog owners is significantly bigger than that of the non-dog owners.

- a. Obtain the data from Section 7.4 in the Coursework area. The csv file is called "Exam1DogData.csv".

(b. continued on next page to utilize space more efficiently and follow one page rule)

- b. (10 points) Test (if possible) to see if the mean of the happiness scores of the dog owners is significantly greater than that of the non-dog owners. Test at the alpha = .01 level of significance.

The data for both groups is not normally distributed – it is heavily right skewed as seen via box plots and histograms:



Even after multiple transformations, the dog data distribution violates the rules of normality (3). Given our sample sizes,  $n > 120$  in both groups, and similar standard deviations (2), however, it would be acceptable to utilize t-tests. With extremely large sample sizes, the tenants of the central limit theorem make the two sample t-tests robust to departures from normality in this study. We will go ahead with a two-sample t-test on the original data:

*Hypothesis:*

$$H_0 : \mu_{dog} = \mu_{nodog}$$

$$H_A : \mu_{dog} > \mu_{nodog}$$

The one sided t critical value on 3972 degrees of freedom at an alpha level of 0.01 is 2.327.

The one sided test statistic (t-statistic) value on 3972 degrees of freedom and an alpha of 0.01 is 23.45. The size of the test statistic indicates we will likely have a small p value. Utilizing the pooled method for the standard deviation assuming equal variances due to a lack of variation in the original data visually (2), the p value is statistically significant at  $p < 0.0001$ .

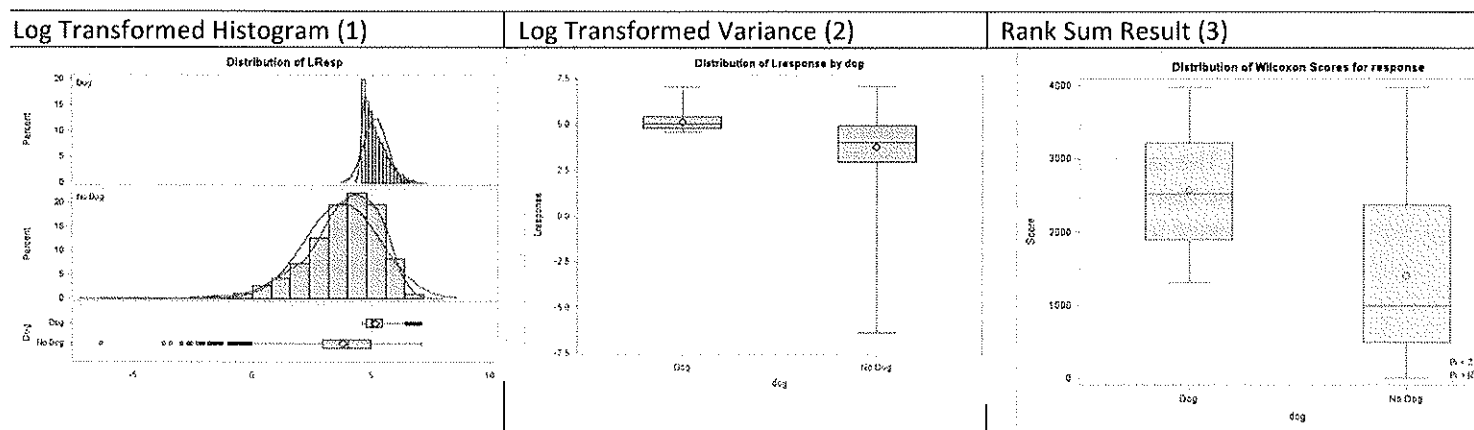
Based on a significant p value of  $< 0.0001$ , we would reject the null hypothesis that the happiness rating of dog owners is the same as the happiness rating of non dog owners.

*Conclusion:*

There is evidence to show ( $p < 0.0001$ ) that dog owners are happier than non-dog owners based on a sample of 3974 people. Dog owners, on average, tend to be 96.81 points happier than non-dog owners according to this study.

- c. (10 points) Test (if possible) to see if the median of the happiness scores of the dog owners is significantly greater than that of the non-dog owners. Test at the alpha = .01 level of significance.

To test the median of the happiness score between the two groups, we will use the log transformed values, which, when back transformed, represent an inference on medians. The log-transformed data still do not present a normal distribution for the dog group, however, given our sample sizes are extremely large, the central limit theorem will allow us to use parametric tests. We can also control for a lack of similar variances between groups (2) with a Welch's T-Test:



Hypothesis:

$$H_0 : \text{Median}_{\text{dog}} = \text{Median}_{\text{nodog}}$$

$$H_A : \text{Median}_{\text{dog}} > \text{Median}_{\text{nodog}}$$

The one-sided t critical value is 2.327.

The one-sided test statistic (t-statistic) value on 2332 degrees of freedom, an alpha of 0.01 and using a log-transformed data set is 36.39. We use a Welch's two-sample t test with a satterthwaite adjustment because there is evidence of unequal variances visually (2) between the two groups when they are log transformed. The size of the test statistic indicates we will likely have a small p value.

Based on a significant p value of <0.0001, we reject the null that the happiness rating of dog owners is the same as the happiness rating of non dog owners. There is a difference of  $e^{1.369}$  between the dog and non-dog groups. To further prove our conclusion and ensure our assumptions are accurate, we run a rank sum (Wilcoxon) permutation test between the two groups and receive a significant one-sided p of <0.0001.

Conclusion:

There is evidence to show ( $p < 0.0001$ ) that dog owners are happier than non-dog owners based on a sample of happiness ratings in 3974 people. On average, dog owners have 3.93 times the happiness points than non-dog owners.



- d. (5 points) Explain how you could use a permutation test to test if the population median of the happiness scores of dog owners is significantly greater than that of the non-dog owners (Do not do any calculations to answer this part).

A permutation test can be used to test the happiness between dog and non-dog owners by permuting values between groups N number of times. Due to the size of each group in the dog study, we would run an appropriate amount of samples to prevent time out. By randomizing the responses between groups, obtaining a test statistic (t-statistic or difference in means) and building a null distribution, we can find the p value of a permutation distribution by dividing the number of permutations resulting in a test statistic greater than our observed test statistic by the total number of permutations. *good*

- e. (5 points) Which analysis do you feel is more appropriate and why?

A parametric test such as the two-sample t-test is acceptable due to the tenants of the central limit theorem. The sample sizes in the dog study are so large we have nothing to worry about when running a parametric test. Further, the variances between the two groups do not differ significantly on the original data, allowing us to use a standard t-test. That being said, a permutation test should at least confirm the results because the dog owner distribution, both log transformed and original, is not normal.

8. (2 points) You are more than halfway done! Take a break and check out this website:

[http://en.wikipedia.org/wiki/William\\_Sealy\\_Gosset](http://en.wikipedia.org/wiki/William_Sealy_Gosset)

List one interesting thing about the man who discovered the Student t distribution. Do not spend a lot of time on this question ...your answer should be a very short sentence!

William Sealy Gosset received mathematics assistance from Karl Pearson, who is responsible for Pearson's R, among other many mathematical and statistical advances. Pretty cool!

9. (10 points) For this problem you will need to use the cityrate.csv file located in Section 7.4. We are analyzing the interest of auto loans in 5 different cities: Chicago, Dallas, LA, NY, and Phoenix. We want to investigate if there is a difference in mean interest rate between the north and the south. Let Chicago and NY represent the north and let Dallas, LA and Phoenix represent the South.

Use a contrast to test the claim at the  $\alpha = .05$  level of significance that the north has a different mean auto interest rate than the south. Be sure and clearly state  $H_0$  and  $H_a$ . You may describe the  $H_0$  and  $H_a$  with respect to  $\mu_i$ 's or  $\gamma$  or show it both ways. Perform a complete analysis: 1) State the problem 2) Address the assumptions. 3) Conduct the Test 4) Clearly state the conclusion in the context of the problem. Also, provide the SAS proc glm statement for this problem.

### 1) Hypothesis:

Given a sample of auto interest rates for Chicago, Los Angeles, Phoenix, Dallas, and New York, we are investigating whether there is a difference in mean rates between auto loans in the northern and southern United States.

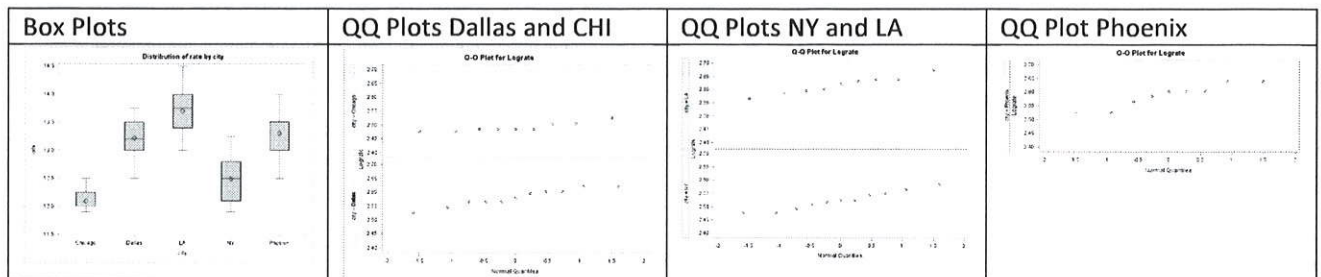
Formal Hypothesis:

$$H_0 : \mu_{\text{NorthRates}} = \mu_{\text{SouthRates}}$$

$$H_A : \mu_{\text{NorthRates}} \neq \mu_{\text{SouthRates}}$$

### 2) Assumptions:

Based on data exploration, our sample sizes are between 9 and 11 per group, roughly normally distributed and the differences in variances between groups are not significant:



Because our sample sizes are between 9 and 11 per group, the distributions are not too far off from normality, and we assume equal variances visually, we will proceed with a contrast of north and south interest rates.

### 3) Conduct the Test

Based on a contrast of north and south regions, we find a large F-value of 80.29 and a significant p value of  $p < 0.0001$ :

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Contrasts for North v South	1	14.62407285	14.62407285	80.29	<.0001

### 4) Establish a Conclusion:

There is evidence to show that auto loan interest rates are lower in the northern regions of the United States compared to the south. Indeed, rates in the north are 1.08 points lower than in the south based on an analysis of mean differences.

```
PROC GLM data = city alpha = 0.05;
  CLASS city;
  MODEL rate = city;
  CONTRAST 'Contrasts for North v South'
    city 1.5 -1 -1 1.5 -1;
RUN;
```

*- difference*

10. Still using the city auto interest rate data, we now want to simply compare Dallas and Chicago. Specifically we would like to test if Dallas has a different mean auto interest rate than Chicago.

- a. (5 points) We of course would like to perform the most powerful test available. Describe whether you would use a simple two-sample t-test using only the data for the two cities or a contrast to compare these two means and why.

Because there are five groups in our study, we should utilize the contrast to compare the two means. The more groups we include, the larger the difference we should expect in the largest and smallest groups. When selecting one comparison out of >2 comparisons, we should include the same statistical measure of uncertainty across all comparisons in order to estimate the population variance more accurately than a two sample t-test, which will not account for variance in the other three groups.

- b. (10 points) Now test the same claim using a contrast by hand. You do not need to actually write it with your hand ... but clearly show the calculations you made to carry out the contrast (typing the equations.) You may skip the assumptions checking and simply show your work in finding the 'g', SE(g), t-statistic and p-value. And of course write a short but complete conclusion.

$$\bar{x}_{Dallas} = 13.2136$$

$$\bar{x}_{Chicago} = 12.0889$$

$$g = -1(12.0889) + 1(13.2136) = 1.1247$$

$$\sqrt{MS_{Error}} = \sqrt{.1822} = 0.4268$$

$$SE(g) = .4268\sqrt{(1/9) + (1/11)} = 0.1918$$

$$t_{critical}(44, d.f., \alpha=.05) = 2.0154$$

$$t = 1.1247 / .1918 = 5.86$$

$$p = < 0.0001$$

#### Conclusion:

There is evidence to show that the mean auto loan interest rates are different between Dallas and Chicago. Based on a contrast between the two cities, auto interest rates are 1.1 percentage points higher in Dallas than in Chicago.

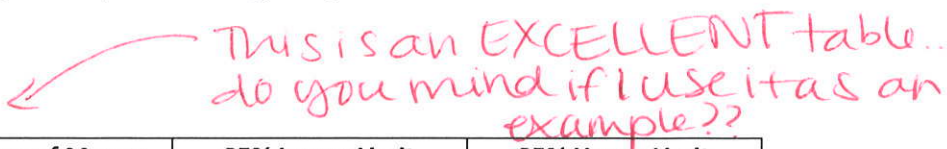


11. (10 points) Let's take a step back now and pretend we had no idea going into this analysis which pairs of cities might be different; so we wish to test all the pairs and see which ones are statistically significant. Use confidence intervals or hypothesis tests to determine which pairs of cities are statistically different. Be sure and address why you chose the methods you chose and defend (if any) assumptions you needed to utilize those methods.

For unplanned comparisons where we test all possible pairwise comparisons, we use the Tukey test. The Tukey test is better for all pairwise comparisons than the Bonferroni method because it is less conservative and controls type II error rate more effectively. Further, Tukey will give us a narrower confidence interval for all pairwise comparisons than the Bonferroni method. The Dunnett's test does not apply, as we do not identify a control in our study. The LSD test is an unadjusted Bonferroni test. Therefore, we move forward with the Tukey method.

Based on the Tukey's HSD test at  $\alpha = .05$ , the following six significant differences and their confidence intervals are observed:

Familywise Comparisons:



Comparison	Difference of Means	95% Lower Limit	95% Upper Limit
LA – NY	1.2091	0.6635	1.7547
LA – Chicago	1.6111	1.0389	2.1833
Phoenix – NY	0.8158	0.2702	1.3613
Phoenix - Chicago	1.2178	0.6456	1.7900
Dallas – NY	0.7227	0.2051	1.2403
Dallas – Chicago	1.1247	0.5792	1.6703

We assume that we have no pre-planned comparisons going into this study, therefore we test all pairwise comparisons ( $n(n-1) = 5(5-1) = 10$  comparisons) with the Tukey method. Further, our acceptance of normality and variance assumptions allows us to perform multiple pairwise comparisons with the Tukey method.