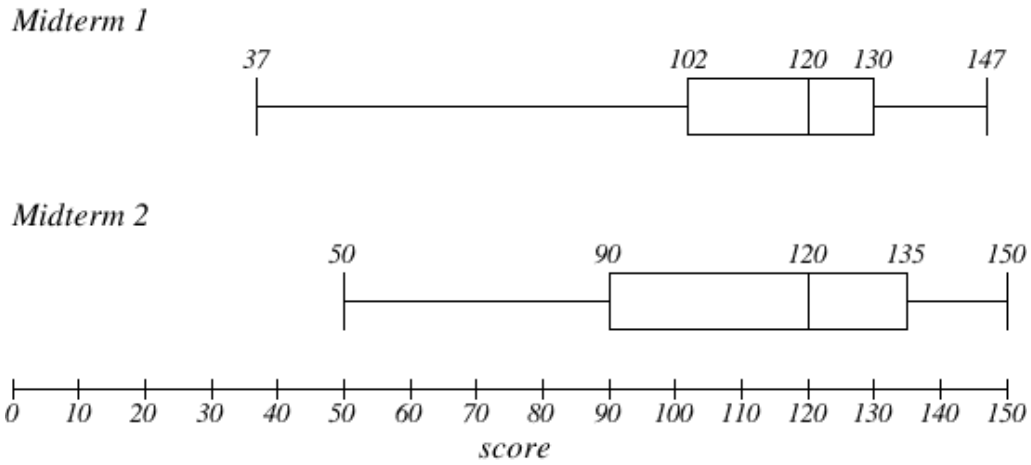
MSDS 6371 Exam 1 SOLUTIONS

Answer the questions that follow to the best of your ability. On questions that require 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 1,2,4,5,6 you may simply highlight the correct answer or make it a different color.

Use the boxplots below to answer questions 1 and 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. Section A
   b. Section B
   c. Both sections 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
   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 result is "statistically significant" if the probability that it would have occurred by chance (pvalue) is very small. In this case, the researcher will assume that the null hypothesis is not true and find in favor of the alternative hypothesis. (The crucial question is "How small?." This value is set by the user and is called the "significance level".)

4. (4 points) An agricultural researcher plants 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 m and that a 95% confidence interval for m is found to be 150 ± 3.29. Which of the following is true?

Either

a. A test of the hypotheses H0: $\mu = 160$, Ha: $\mu \neq 160$ would be rejected at the 0.01 level.

b. A test of the hypotheses H0: $\mu = 150$, Ha: $\mu \neq 150$ would be rejected at the 0.05 level.

c. A test of the hypotheses H0: $\mu = 150$, Ha: $\mu > 150$ would be rejected at the 0.05 level.

d. A test of the hypotheses H0: $\mu = 160$, Ha: $\mu \neq 160$ would be rejected at the 0.05 level.

Remember that the confidence interval contains plausible values of the mean. This interval is centered on 150 meaning that 150 is the most plausible value of the mean given the data. For this reason, if our null hypothesis is Ho: mu = 150, then we would certainly fail to reject this hypothesis. Since the interval given (146.71, 153.29) does not contain 160, 160 is considered a plausible value of the mean and a test of mu = 160 would be rejected. Therefore Answer D is the answer.

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? **P(Type II Error) = 1 – .8 = .2**

b. What is the value of the probability of Type I error? **P(Type I Error) = .05**

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

If the standard deviation decreases that means we have more accuracy. And if we have more accuracy we don't need as large a sample size to achieve the same power. For this reason the sample size (ntotal) will decrease.

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.

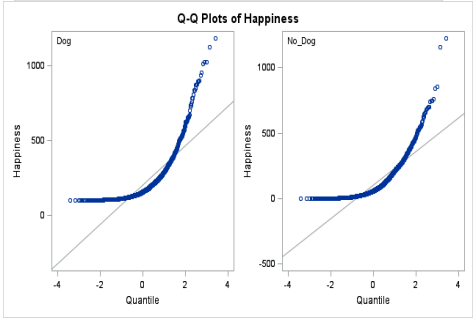
No… the Ho is either correct or it is not correct. There is a 3.2% chance that *if the Ho is correct*, we would see what we saw or more extreme by random chance (.032 is a p-value))

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 μ and σ. Finally finally, remember that a "Test" includes addressing the assumptions, running the 5 steps 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. (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.
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.
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).
e. (5 points) Which analysis do you feel is more appropriate and why?

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 SAS System

The TTEST Procedure

Variable: Happiness

Dog	N	Mean	Std Dev	Std Err	Minimum	Maximum
Dog	1996	200.8	132.0	2.9538	100.0	1183.0
No_Dog	1978	104.0	128.3	2.8839	0.00172	1223.8
Diff (1-2)		96.8054	130.1	4.1287		

Dog	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
Dog		200.8	193.2	208.4	132.0	126.8	137.6
No_Dog		104.0	96.5266	111.4	128.3	123.2	133.7
Diff (1-2)	Pooled	96.8054	86.1654	107.4	130.1	126.5	134.0
Diff (1-2)	Satterthwaite	96.8054	86.1668	107.4			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3972	23.45	<.0001
Satterthwaite	Unequal	3970.5	23.45	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	1995	1977	1.06	0.2049

Ans:

Assumptions:

Normality: According to the histograms and the QQ plots (above), there is strong evidence that the populations are not normal. Since the sample size is large, if one or both population are non-normal, the large sample size is sufficient to invoke normal distributions of the sample mean through the CLT. From the box plots, we can see the presence of some outliers. The large sample size should mitigate the effects of these outliers again through the CLT.

Independence: From the question, we know that the data consist of 3974 happiness values recorded from a random sample of 3974 people from Missouri. Hence, we will assume these observations are i.i.d.

Identical standard deviation: From the histograms and box plots, there is minimal evidence to suggest the standard deviations are different. We will proceed under the assumption of equal standard deviation. (A Brown-Forsythe test could be conducted here as well for secondary evidence.)

Basing on above, it is appropriate to pool the variances; Since the t-test is robust to departures from normality when the sample size is large, we will choose the pooled t-test account for the same standard deviation to test the means of the happiness scores of the dog owners and the non-dog owners.

Test: Five steps:

1: $H_0: u_d \leq u_{nd}$

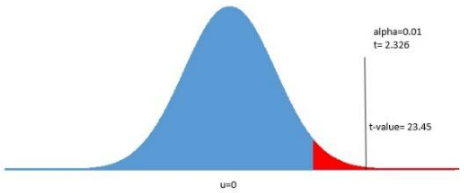
$H_1: u_d > u_{nd}$

2: $t_{0.01,3972} = 2.326$

3: t-value=23.45

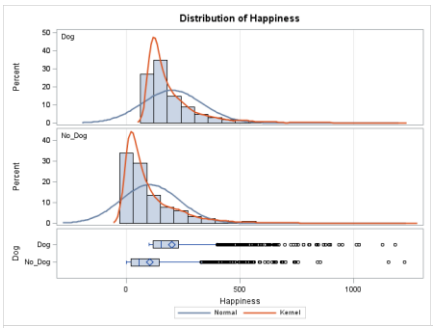
4: p-value<0.0005 $< \alpha = 0.01$

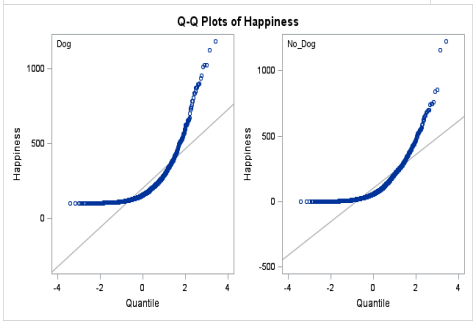
5: reject H_0



Conclusion: There is sufficient evidence at alpha=0.01 level of significance (p-value<0.0005 from the pooled t-test) to suggest that the mean happiness score of the dog owners is greater than the mean happiness score of the non-dog owners. Our best estimation of the difference in means is 96.81 points. A 98% confidence interval of the mean happiness score is (87.19, 106.41) points.

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.





The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Happiness Classified by Variable Dog

Dog	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
No_Dog	1978	2783132.50	3931275.0	36163.5749	1407.04373
Dog	1996	5115192.50	3967050.0	36163.5749	2562.72169

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	2783132.5000
Normal Approximation	
Z	-31.7486
One-Sided Pr < Z	<.0001
Two-Sided Pr > |Z|	<.0001
t Approximation	
One-Sided Pr < Z	<.0001
Two-Sided Pr > |Z|	<.0001
Z includes a continuity correction of 0.5.	

Kruskal-Wallis Test	
Chi-Square	1007.9729
DF	1
Pr > Chi-Square	<.0001

Ans:

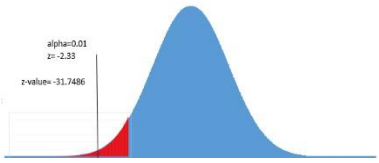
Assumption:

The data are ordinal and can thus be ranked. The Rank-Sum Test is distribution free therefor no distribution based assumptions exist. We will assume the data are independent.

Test: Five steps:

1: $H_0: Median_{nd} \geq Median_d$

$H_1: Median_{nd} < Median_d$



2: $z_{0.01} = -2.33$

3: z-value= -31.7486

4: p-value$< 0.0001 \quad < \alpha = 0.01$

5: reject H_0

Conclusion: There is sufficient evidence at alpha=0.01 level of significance (p-value< 0.0001 from the rank sum test) to suggest that the median happiness score of the dog owners is greater than the median happiness scores of the non-dog owners. Our best estimation of this difference the difference of the sample medians: 97.63 points.

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).

In order to conduct a permutation test one would we need to first record the difference in the sample medians (Dog – No Dog) between the two groups. Next the labels "Dog" and " No Dog" should be randomly assigned to the data and the difference of the sample medians should again be recorded. This process should be repeated many times (up to the user) and each difference of sample medians should be recorded. Finally, the percentage of the difference of medians that are as extreme or more extreme than the difference that was observed is recorded as the pvalue. The conclusion should be made with respect to this pvalue as usual.

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

Since the raw appear to be right skewed, the median may be considered a better measure of center than the mean. For this reason I would choose the rank some test over the ttest.

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

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

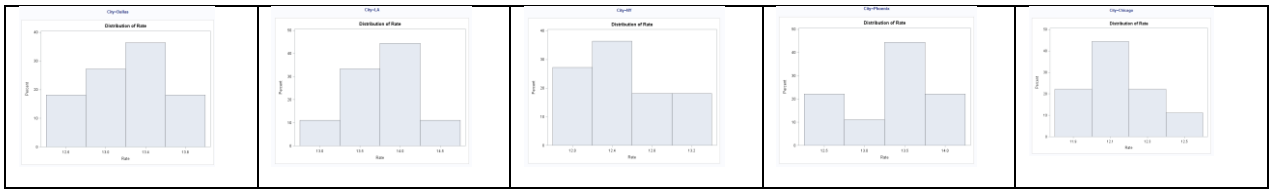
   List one interesting thing about the man who discovered the Student t distribution. Do not spend a lot of time on this question … answer should be a very short sentence!

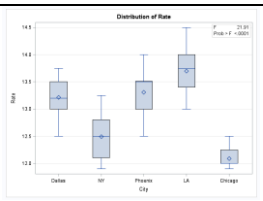
   He worked for the Guinness Brewery and discovered the t-distribution … aka the "Student" t- distribution.

9. For this problem you will need to use the cityrate.csv file. We are analyzing the interest of auto loans in 5 different cities: Chicago, Dallas, LA, NY, and Phoenix.
   a. (10 pts) We want to investigate if there is a difference in mean interest rate between the north and the south. Let Chicago and NY (New York) 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 Ho and Ha. You may describe the Ho and Ha with respect to μ_i's or γ 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.) State the Problem: We want to test the claim at the alpha = .05 level of significance that the north has a different mean auto interest rate than the south.
      2.) Assumptions:
         1. Normality: Judging from the histograms below, there is not significant evidence that the data from each city do not come from a normal distribution. We will assume that the data come from normal distributions.
         2. Equal Standard Deviations: Judging from the histograms and boxplots below, there is not strong evidence that the data come from distributions with different standard deviations. Since the data from New York have a bigger sample standard deviation that those from Chicago the Brown Forsythe test was used for secondary evidence. This test is also consistent with equal standard deviations between cities (pvalue = .2207). Therefore, we will proceed under the assumption that the data come from distributions with the same standard deviations.
         3. We will assume the data are independent.

Dallas	LA	New York	Phoenix	Chicago





Brown and Forsythe's Test for Homogeneity of Rate Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
City	4	0.4190	0.1048	1.49	0.2207
Error	44	3.0875	0.0702		

3.) Conduct the Test

Test: Five steps:

1: $H_0: \mu_{North} = \mu_{South}$

$H_1: \mu_{North} \neq \mu_{South}$

2: $F_{1,44} = 4.06$

3: F value= 80.29

4: p-value$< 0.0001 \quad < \alpha = 0.05$

5: reject H_0

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

4.) The evidence suggests (alpha = .05, pvalue < .0001) the mean interest rate of the North is different than that of the South. The best estimate of this difference is the difference in the sample means: 1.08. 95% confidence interval = (.8136, 1.35). These last statistics were from proc ttest.

5.) SAS CODE:

```
proc glm data = cityrate order = data;
class city;
model rate = city;
means city / hovtest = bf;
contrast "North Vs. South" city  2 -3 2 2 -3;
run;
```

```
proc ttest data = cityrate;
class Region;
var rate;
run;
```

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 pts) 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.

    If we can assume the standard deviations are equal, we should use a contrast. The contrast uses the pooled estimate of the standard deviation: s_p. This estimate uses all of the groups (cities) thus increasing our degrees of freedom and therefore the power.

b. (10 pts) 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 calculation you made to carry out the contrast (typing the equations.) You may skip the assumptions and the 5 steps 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.

Level of City	N	Rate	
		Mean	Std Dev
Chicago	9	12.0888889	0.20121161
Dallas	11	13.2136364	0.40440754
LA	9	13.7000000	0.45893899
NY	11	12.4909091	0.43521155
Phoenix	9	13.3066667	0.55634971

$$g = (1)(12.089) + (-1)(13.214) = -1.125$$

(The pooled SD (sp) can be found by taking the square root of the MSE from the corresponding ANOVA table.

ANOVA TABLE FROM FULL MODEL ABOVE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	15.96389219	3.99097305	21.91	<.0001
Error	44	8.01463434	0.18215078		
Corrected Total	48	23.97852653			

$$SE(g) = \sqrt{.1822} * \sqrt{\frac{(1)^2}{9} + \frac{(-1)^2}{11}} = .19185$$

$$t = \frac{g}{SE(g)} = \frac{-1.125}{.19185} = -5.86$$

$$Pvalue = P(t_{44} < -5.86) < .0001$$

The evidence suggests at the alpha = .05 level of significance (pvalue < .0001) that the mean auto interest rate of Dallas and Chicago are different. It should be noted that we had to make a strong assumption of the equality of the standard deviations in this test and suggest that further analysis (bigger sample sizes) may be needed as to the distributions of the auto interest rates in these cities.

As an extra note: remember that you can run the contrast statement in SAS and get the following table below:

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Chicago v. Dallas	1	6.26203157	6.26203157	34.38	<.0001

Notice that this is an F test: yet it is equivalent to our results. Theory tells us the $t^2 \sim F_{1,df.}$ We can see this in the current example by noticing that $t^2 = (-5.86)^2 = 34.38$. Also, the pvalue that was calculated in the table comes from a $F_{1,44}$ distribution. Cool right?

11. (10 pts) 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.

Step 1: State the problem: Identify the pairs of cities that have statistically different mean auto interest rates.
Step 2: Assumptions: The Tukey-Kramer multiple comparison procedure is simply a modified t-test and thus has all the assumptions that come with the ttest. We have established above that we can assume, with caution, that the data come from normal distributions with equal standard deviation. In addition we have assumed that the data are independent. Therefore, we will use the Tukey-Kramer procedure to identify, at the 95% confidence level, those cities that have different mean auto interest rates.
Step 3: The Test

Tukey's Studentized Range (HSD) Test for Rate

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	44
Error Mean Square	0.182151
Critical Value of Studentized Range	4.02217

Comparisons significant at the 0.05 level are indicated by ***.

City Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
LA - Phoenix	0.3933	-0.1789	0.9655	
LA - Dallas	0.4864	-0.0592	1.0319	
LA - NY	1.2091	0.6635	1.7547	***
LA - Chicago	1.6111	1.0389	2.1833	***
Phoenix - LA	-0.3933	-0.9655	0.1789	
Phoenix - Dallas	0.0930	-0.4525	0.6386	
Phoenix - NY	0.8158	0.2702	1.3613	***
Phoenix - Chicago	1.2178	0.6456	1.7900	***
Dallas - LA	-0.4864	-1.0319	0.0592	
Dallas - Phoenix	-0.0930	-0.6386	0.4525	
Dallas - NY	0.7227	0.2051	1.2403	***
Dallas - Chicago	1.1247	0.5792	1.6703	***
NY - LA	-1.2091	-1.7547	-0.6635	***
NY - Phoenix	-0.8158	-1.3613	-0.2702	***
NY - Dallas	-0.7227	-1.2403	-0.2051	***
NY - Chicago	0.4020	-0.1436	0.9476	
Chicago - LA	-1.6111	-2.1833	-1.0389	***
Chicago - Phoenix	-1.2178	-1.7900	-0.6456	***
Chicago - Dallas	-1.1247	-1.6703	-0.5792	***
Chicago - NY	-0.4020	-0.9476	0.1436	

Step 4: Conclusion: According to the Tukey-Kramer adjusted confidence intervals above (from SAS proc GLM), the pairs of cities that have significantly different mean auto interest rates are: LA and New York, LA and Chicago, Phoenix and New York, Phoenix and Chicago, Dallas and New York, Dallas and Chicago