Stat 302, Assignment 2, Due Thursday February 18, 2016 at 4:30pm

There are 4 multi-part questions and a set of questions based on 5 pages of reading.
Please see the Stats Workshop for help, or see me in office hours (Tue 1-2, Thur 3:30-4:30), or e-mail me at jackd@sfu.ca

Explanations and relevant computer output should be included, but plots do not need to be.
Necessary R code is included for every question, after the reading questions.
There are also four practice problems, mostly surrounding review material.

Total	/ 64	Name _____
Q1	/ 7	
Q2	/ 14	
Q3	/ 10	Student Number _____
Q4	/ 16	
Reading	/17	

1) Consider the dataset A2_npk.csv , which is a dataset of yield across several different factors for some industrial process.

1a, 2pts) Check this data against unequal variance and imbalance. Are there any problems? How do you know?

1b, 1pt) Using the p-value, use ANOVA test the hypothesis of no difference between the four groups.

1c, 2pts) Using the obtained F score, test whether there is any difference between the four groups. Use the $\alpha = 0.05$ significance level.

1d, 2pts) How much of the variance in **yield** is explained by the different levels of **block**? Show your work.

2) Consider the dataset A2_Toothgrowth.csv, which measures the tooth growth of Guinea pigs when given two different kinds of supplement (Vitamin C and Orange Juice) at three different dosage levels (0.5, 1.0, ans 2.0 mg)

2a, 2pts) Conduct a one-way ANOVA using **supplement** as a grouping variable and **growth** as a response variable. Does the response vary between different groups? Use the $\alpha = 0.05$ significance level.

2b, 2pts) Conduct a one-way ANOVA using **dosage** as a grouping variable and **growth** as a response variable. Does the response vary between different groups? Use the $\alpha = 0.05$ significance level.

2c, 3pts) Conduct a two-way ANOVA using the same response both grouping variables from 2a and 2b. Using the p-values, test whether the response changes across different types of **supplement** and of **dosage** changes, holding each other constant. Use the $\alpha = 0.05$ significance level.

2d, 3pts) Using the obtained F scores, conduct the same two tests as you did in 2c.

2e, 2pts) Explain the apparent paradox between the results of 2a and 2c. Why does **supplement** appear to be a non-significant (or marginally significant) factor in one answer, and strongly significant in the other?

2f, 2pts) How much of the variance in **growth** is explained by **supplement** and by **dosage**? Show your work.

---

3) Consider the dataset A2_chickwgt.csv, which measures the weight of chickens that were given six different kinds of feed supplement.

3a, 1 pt) Conduct a one-way ANOVA using **feed** as a grouping variable and **weight** as a response variable. Does the response vary between different groups? Use the $\alpha = 0.05$ significance level.

3b, 2 pts) If we were to compare each pair of means individually using a t-test and a Bonferroni correction with a family/experiment-wise α of 0.05, what is the alpha used for each pair to determine significance?

3c, 4 pts) Using Tukey's Honestly Significant Difference analysis, which pairs of means show a significant difference? Use a family/experiment-wise α of 0.05.

3d, 3 pts) Using the results from 3c, arrange the groups into clusters. Label the group means that aren't honestly different from the highest

---

4) Consider the dataset A4_viral.csv, which is loosely based on real data of HIV viral load in blood in response to three different treatments and a control.

4a 2 pts) Use **load** as a response and **trt** as a grouping variable. Check the ANOVA against unequal variance and imbalance. Are there any problems? How do you know?

4b 1 pt) From a boxplot, would you say there is a difference in viral load between the groups? Briefly explain.

4c 2 pts) Ignore any problems and conduct a one-way ANOVA on **load** as a response to **trt**. Using the p-value, test the hypothesis of no difference between the groups. Does this match your expectations from the boxplot?

4d 3 pts) Using Tukey's Honestly Significant Difference analysis, which pairs of means, if any, show a significant difference? Use a family/experiment-wise α of 0.05, and list p-values of significant differences. Given your answer in 4c, is this surprizing?

4e 2 pts) Now use **logload** (the log-transform of **load**) as a response and **trt** as a grouping variable. Check this new ANOVA against unequal variance and imbalance. Comment breifly.

4f 1 pt) From a boxplot, would you say there is a difference in **log(viral load)** between the groups? Briefly explain.

4g 2 pts) Repeat the one-way ANOVA for this new, log-transformed response. Test the hypothesis of equal means again. Does this match your expectations from the new boxplot?

4h 3 pts) Repeat the Tukey's HSD with the log-transformed response. Use a family/experiment-wise α of 0.05.

Reading questions)
These questions pertain to "The Insignificance of Significance Testing", by Neville Nicholls.
The Annals of Applied Statistics
May 2000, Vol. 81, No. 5, 981-985

The answer to R1 appears first in the text, and so on. Every question can be answered in 25 words or fewer.

R1, 2 pts) Give an example of a data set that could have physical significance, but not statistical significance.

R2, 2 pts) Cohen mentions some probability when talking about an n=50 sample from a population with population correlation of $\rho = 0.30$. What is the name for this probability? (The name isn't in the paper, it IS in our notes on hypothesis testing)

R3, 2 pts) The 1979-95 data used to calculate climate trends has no uncertainty from sampling, How is this possible?

R4, 3 pts) What deviations from a well-behaved distribution could affect the correlation between SOI and snowfall?

R5, 3 pts) What are three alternatives to null hypothesis testing?

R6, 2 pts) What is another term for the "repeated investigations" issue? That is, when a so-called insignificant result is left unpublished until someone repeats it and finds significance simply by chance? (Again, the term is in our notes, think Tukey)

R7, 3 pts) A permutation test produces a value that works very similarly to a p-value. However, a permutation test has one major advantage over a classic hypothesis test (and confidence intervals). What is it? (Hint: Spearman correlation and the median have this same advantage).

```
########### EXAMPLE CODE QUESTION 1
## Get the dataset npk, which is embedded into R
Q1 = npk

### Forces 'block' to be interpreted as a categorical variable,
### not a numerical one.
Q1$block = as.factor(Q1$block)

##  One-way ANOVA on block level
mod = lm(yield ~ block, data=Q1)
anova(mod)

## Get the critical F value for 5 grouping df, and 18 residual df
qf(.95, df1=5, df2=18)



########### EXAMPLE CODE QUESTION 2
## Get the dataset ToothGrowth, which is embedded into R
Q2 = ToothGrowth
Q2$dose = as.factor(Q2$dose) # Treat the dosage like a categorical
variable

## One-way ANOVA on supplement type
mod1 = lm(len ~ supp, data=Q2)
anova(mod1)

## One-way ANOVA on dosage level
mod2 = lm(len ~ dose, data=Q2)
anova(mod2)

## Get the critical F values for models 1 and 2
qf(.95, df1=1, df2=58)
qf(.95, df1=2, df2=57)

## Two-way ANOVA on both dosage AND supplement
mod12 = lm(len ~ supp + dose, data=Q2)
anova(mod12)
```

```
########### EXAMPLE CODE QUESTION 3
## Get the dataset chickwts, which is embedded into R
Q3 = chickwts

## One-way ANOVA on feed type
mod = lm(weight ~ feed, data=Q3)
anova(mod)

## Tukey's HSD Analysis
mod = aov(weight ~ feed, data=Q3)
TukeyHSD(mod, ordered=TRUE)



########### EXAMPLE CODE QUESTION 4
### Get the dataset viral, which is made up for this dataset
Q4 = read.csv("A2_viral.csv")

### Get the mean, standard deviation, and N from each treatment group
by(Q4$load, Q4$trt, mean)
by(Q4$load, Q4$trt, sd)
by(Q4$load, Q4$trt, length)

### Produce a boxplot of load by treatment group
boxplot(load ~ trt, data=Q4)

### One-way ANOVA of load by treatment group
mod = lm(load ~ trt, data=Q4)
anova(mod)

## Tukey's HSD Analysis
mod = aov(load ~ trt, data=Q4)
TukeyHSD(mod, ordered = TRUE)


### Now repeat the same things with logload instead of load
```