Stat 302, Assignment 1, Due Thursday January 28, 2016 at 4:30pm

(This due date overrides the syllabus due date of Jan 26).

There are six multi-part questions and a set of questions based on 11 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 /65
Q1 /6
Q2 /7
Q3 /8
Q4 /4
Q5 /10
Q6 /12
Reading /18

Name _____

Student Number _____

1) Consider the different types of data. Specifically numeric and categorical.
[6 points, 2 per question.]

1a) Name two discrete numeric variables, and give a possible value for each.

1b) Name two continuous numeric variables, and give a possible value for each.

1c) Name two categorical variables, and give a possible set of values for each.

---

2) Consider this dataset from the a collection of measurements of iris setosa flowers. (A1setosa.csv)
[7 points total. Do not include plots.]

2a) Describe the five variables of the dataset. Are they numeric or categorical? Are they discrete or continuous? [2 pts]

2b) Find the mean and median of the Sepal length. [1 pt]

2c) Create a histogram of the distribution of the first variable (Sepal length). [4 pts, do not include histogram]
Use the histogram and the mean and median to describe this distribution.
Is it unimodal (one peak), or multimodal (multible peaks)?
Is it left/negatively skewed (extreme values on left/lower), or right/positively skewed (extreme values on right/higher)?
Explain how you know this from the mean and the median.

3) Consider this dataset of the spending habits of 100 families. Specifically, their food spending per month and their clothing spending per year. (A1spend.csv)
[8 points, 2 per question. Do not include plots.]

3a) Find the Pearson correlation coefficient. Test the hypothesis that the parameter rho = 0 at the 0.01 level?

3b) How much of the variation in clothing spending can be explained by food spending?

3c) Verify the statistical significance of the correlation with a two-sided t-test. Report the t-score, degrees of freedom and p-value.

3d) Produce a scatterplot to better see the relation. Is there any trend in the plot that could be a problem?

---

4) The correlation between two abstract variables. X and Y, is r = +0.76
[4 points, 1 per question.]

4a) What proportion of the variance in Y is explained by X?

4b) If X and Y were switched, what would the correlation become?

4c) If Y were multiplied by 10, what would the correlation become?

4d) If Y were multiplied by -10, what would the correlation become?

5) Consider the gross domestic product (GDP, a measure of the mean income of the citizens in a country), and life
expectancy (at birth) in the countries of the world in 2003. (A1countries).
[10 points, 2 per question. Do not include plots.]

5a) Find the Pearson correlation coefficient. Is it significant at the 0.01 level?

5b) How strong is this correlation?
Does it fully describe the strength of the relationship between GDP and life expectancy?

5c) Produce a scatterplot of Life expectancy(y) over GDP(x) and describe it to back up your answer in 5b.

5d) What else might better capture the strength of this relationship? Why?

5e) Confirm you answer in 5d by finding a different correlation between GDP and life expectancy, and by find the Pearson correlation between Log10GDP and life expectancy.

---

6) Consider the dataset of standings in the National Hockey for 2011-12 regular season, in "A1nhl.csv"
Create a linear regression of wins (W) in response to goals scored against (GA).
To clarify, wins is the 'y' variable, and goals against is the 'x' variable.
[12 points, 2 per question. Do not include plots.]

6a) Determine the least-squares estimates of the intercept and slope parameters.

6b) Test the null hypothesis of zero slope. Interpret the result.

6c) Find a 95% confidence interval of the slope of wins (W) over goals against (GA).
Explain your choice for degrees of freedom.
Explain your choice for quantile (hint: 95% refers to the MIDDLE 95% of the distribution)
Interpret the confidence interval.

6d) The value -0.2 is inside the 95% confidence interval for the slope in 6c. Using only this information, can you tell whether -0.2 is inside the 99% confidence interval or not? Explain.

6e) Plot the residuals of this regression against GA. Are there any potential problems visible? Explain.

6f) Plot the Cook's Distance for this regression. Are there any potential problems visible? Explain.

Reading questions)
These questions pertain to the first 3 sections of "For Objective Causal Inference, Design Trumps Analysis", by Donald B. Rubin.
The Annals of Applied Statistics
2008, Vol. 2, No. 3, 808–840 (Stop at Page 818)

The answer to R1 appears first in the text, and so on. [18 points, 3 per question.]

R1) Randomized experiments make causal inference valid because we know the 'scores' "are known from the design of the experiment". What are these 'scores'? (Name only)

R2) Rubin is stating that randomized experiments are 'the gold standard' for causal inference. Does that mean causality can be inferred from all randomized experiments, or are some "poorly suited"? If so, give an example.

R3) What are the two design steps that are "absolutely essential" for objective inferences?

R4) In Section 2.1, how is a treatment defined?

R5) What are covariates, "in contrast to" outcome/response variables?

R6) What "self optimizing" behaviour happens when treatments are not assigned randomly by the experimenter?

```
####### Example code for Question 2
## (Anything after a # is treated as a comment in R, not code)

## Opening from a file
## First, go to file --> Change dir.. and change to the directory
where  A1setosa.csv is saved
Q2 = read.csv("A1setosa.csv")


## Look at the first few rows/cases of data
head(Q2)

## Look at the structure of the variables
str(Q2)

## Get a histogram
hist(Q2$Sepal.Length)


## Get the mean and median
mean(Q2$Sepal.Length)
median(Q2$Sepal.Length)

## Get summary information
summary(Q2$Sepal.Length)



####### Example code for Question 3
Q3 = read.csv("A1spend.csv")
head(Q3)

# Find the correlation, and test it
cor(Q3$food, Q3$clothing)
cor.test(Q3$food, Q3$clothing)

# Make a scatterplot
plot(Q3$food, Q3$clothing)
```

```
#### Example code for Question 5
Q5 = read.csv("A1countries.csv")
head(Q5)

# Test the correlation
cor.test(Q5$LifeExp, Q5$GDP)

# Make a scatterplot
plot(Q5$LifeExp, Q5$GDP)

# Take some more correlations
cor(Q5$LifeExp, Q5$GDP, method="??????")
cor(Q5$LifeExp, Q5$Log10GDP, method="pearson")



####### Example code for Question 6
nhl = read.csv("A1nhl.csv")

### Make and summarise the regression
mod = lm(W ~ GA, data=nhl)
summary(mod)


### Get the critical value
#!!! Note that the df and q are deliberately wrong in this line. Part
of the exercise is find the right answer.
qt(.900, df=14)

### Residual Plot
plot( resid(mod) ~ nhl$GA)

### Getting Cook's Distance, a leverage measure
plot(cooks.distance(mod))
```

Practice Problem 1)
When nutrition guides were being made, a large group of healthy people's diets were surveyed. The daily intake of any given nutrient was found to be normally distributed.
The amounts that were two standard deviations below the mean were set as the recommended daily minimums.

a) What proportion of surveyed healthy people met or exceeded this minimum for calcium?

b) Consider three surveyed healthy people. What is the probability that all three are getting enough calcium?

c) Adult men in the survey took in 14 mg/day of zinc, with a standard deviation of 1.5 mg/day, what is the daily recommended minimum of zinc for adult men?

d) All adult men, healthy or otherwise, take a mean of 12mg/day of zinc and a standard deviation of 2 mg/day. What proportion of them get their daily recommended amount of zinc?

Practice Problem 2)

Which are statistics and which are parameters?
Identify the population of interest for every question and identify the sample when applicable.
(Sources are for interest only, everything you need is in the sentence.)

a) "The federal agency said the census data showed that [85.2 %] of Canada's population was under the age of 65
(Source: Canada.com, "Canada still youngest amongst G8 "May 29,2012)

b) 25 courses were randomly selected from all those available in the summer, and an mean of 2.4 required textbooks was found

c)A research organization conducted a survey and found that 45%of respondents felt they were better off economically than before.

d)"Statistics Canada says the caseload in youth courts fell seven per centin 2010-2011, a second straight annual drop.The agency says youth courts handled about 52,900cases last year."
(source: Toronto Star: "Youth Courts Caseload Falling: Statscan. May 28, 2012")

e) “A survey featured in [a recent] report reveals 82 per cent of mothers cite safety concerns as reasons why they restrict outdoor play
[of their children].”
(Source: The Globe and Mail: “Canadian kids get failing grade in physical activity: report.” May 29, 2012)

Practice Problem 3)
Hummus, a food product, comes in packages that claim there is 227 grams of hummus inside the container. In reality not every container has the exactly the same amount of hummus.
The weight of the product has a normal distribution with mean 232 grams, and standard deviation of 4 grams.

a) What is the probability that a randomly selected hummus package is underweight (contains less than the advertised amount)?

b) What is the probability that an average of 4 packages is less than 227 grams?

c) What is the probability that an average of 25 packages is less than 227 grams?

d) Give an upper bound to the chance that an average of 400 packages is less than 227 grams?

e) If you took the average of a larger and larger sample, what would expect to happen to the sample mean of the weights?

Practice Problem 4)

In each situation, would a large value of significance level alpha (more than .05) or a small level of alpha be more appropriate (.05 or less)?

A) A trial for which the penalty is death (null hypothesis is innocence).

B) Determining if you should take extra vitamin C to ward off a cold this morning (null hypothesis is you’re not in danger of getting a cold).

C) A patient might have a disease that would require cutting off a leg. (Null hypothesis is that the leg doesn’t need to be cut off).

D) You have to decide whether to look before crossing the street. (Null hypothesis is that it’s safe to cross without looking).

E) You have an exam in the morning and you’re deciding whether to set the alarm or not. (Null hypothesis is that you’ll wake up in time without an alarm?