STAT 540, Fall 2015

Due on **12:00 A.M., Friday, September 4th**

Homework 1

Homework format for all STAT 540 homework this term: Please label all problems clearly and turn in an organized homework assignment. You don't need to spend hours producing beautifully typeset homework, but you won't get credit if we can't find or read your answer. Unless noted otherwise, turn in the following (as appropriate for the problem).

- Theoretical derivation (when asked for).
- Numerical results **with an explanation of your solution**, written in complete sentences. If computer code is absolutely necessary to provide context here, then include it–nicely formatted–within the solution (otherwise, see below).
- Appropriate graphics. Use informative labels, including titles and axis labels. Try to put multiple plots on the page by using, for example, the R command `par(mfrow=c(2,2))`.
- **Only as necessary**: Final clean computer code used to answer the problem **attached to the end of your homework**. Only include the rare code excerpts without which we wouldn't be able to figure out what you did. Annotate your code. Number and order the code in order of the problems. When in doubt, leave it out; consider that we will probably never read it.
- Some problems will be relatively open-ended, such as "Here are some data. Analyze them and write a report." I will provide further instructions about reports later. They should be self-contained, with suitable EDA, graphs, numerical results, and **scientific interpretation**. No computer code should be included. The report should be concise: "no longer than necessary".

(1) Review of probability theory:
   (a) Imaging extending a string from $(0,0)$, the origin in $\mathbb{R}^2$, to a random point (x,y) in $\mathbb{R}^2$, where $x \sim N(3,1)$ independent of $y \sim N(0,1)$. Use `R` to find the probability that the string will need to be longer than 4 units from $(0,0)$ to (x,y).
   (b) Suppose that z_1, z_2 are i.i.d. $N(0,1)$. Find the distribution of the following random variables.
      (i) $(z_1 - z_2)^2/2$
      (ii) $(z_1 + z_2)/|z_1 - z_2|$

(2) Review of some fundamental statistics.

(a) Consider the variance of the estimated difference between two treatment means when variances are homogeneous, which is denoted by σ^2. To examine this numerically, compute the values of $\text{Var}(\bar{Y}_1 - \bar{Y}_2)$ as a bivariate function of the sample sizes n_1 and n_2. What conclusion you can draw regarding the variance with respect to n_1/n_2?

(b) A study recorded children's cholesterol level and the number of hours of tv watched per day. Children who watched more than two hours of television each day tend to have higher cholesterol levels than children who watched less than two hours of television daily.

(i) Is this an observational or experimental study?

(ii) A newspaper report on this study has the headline: "Television watching bad for your health". Is this headline an appropriate summary of the study? Explain why or why not.

(3) The next three problems give you a short description of a study. For each of them:

i. say whether this is an observational study, an experimental study with non-random assignment of treatments, or a randomized experiment.

ii. For randomized experiments, identify the observational unit and the experimental unit.

(a) An evaluation of anti-smoking treatment effectiveness. 52 smokers in Los Angeles County volunteered for the study. Each was randomly assigned to one of two treatments: the "Spiegel" method and a control. Each smoker was asked to record the number of cigarettes smoked per week. This information was recorded for the first week after initiation of treatments, the second week, and the third week.

(b) An evaluation of the effect of neighborhood poverty on reading proficiency. Census data are used to identify 10 neighborhoods with average income near the poverty line (poor) and 10 neighborhoods with average income more than twice the national average (rich). The school closest to the geographic center of each neighborhood is identified. The reading proficiency of each 8th grade students at that school is measured using a standardized test. The data are the reading scores for each student.

(c) Exactly the same as the previous study, except you are not given data on individual students. You are only given the average reading score for all 8th grade students at the school.

(4) Consider the Tolcua Company example in the textbook (Table 1.1). The file `ch01ta01.txt` is available on the course webpage and contains the data on lot size and work hours, defined as the number of labor hours required to produce the lot. The aim of the example is to investigate the relationship between work hours and lot size. Perform 1-D and 2-D exploratory analysis using `R` with this aim in mind.

(5) Show the MLE estimator of σ^2 on page 24 in lecture 2 satisfies $\mathbb{E}(\hat{\sigma}^2_{MLE}) = (n-2)/n\sigma^2$, which is therefore biased.

(6) Show the properties of fitted regression line in lecture 2, that is for $e_i = y_i - \hat{y}_i$

(a) $\sum_{i=1}^n e_i = 0$,

(b) $\sum_{i=1}^n x_i e_i = 0$,

(c) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.

(7) Textbook problems:

(a) Problems 1.4; 1.5; 1.12

(b) Problem 1.27

(c) Problem 1.32; derive the expression for β_1 in the textbook (1.10a) from the normal equation in (1.9). Perform this also for β_0, and derive the estimator of σ^2 given on page 24 in lecture 2.

(d) Problem 1.37: continued example of Toluca Company.

(e) Problem 2.1 parts (a) and (b); 2.7

Self-taught R: (Do not turn in!)

(1) Install the `car` package if you don't have it. It contains the dataset `Mroz`. See `help(Mroz)`. Produce a histogram of a suitable variable in this dataset.

(2) Use the `density()` function on the same variable and produce a nice plot.

(3) Make a scatterplot of two suitable variables from `Mroz`.

(4) Use `identify()` to add labels to a few of the points in your plot. Make the labels be the case number (i.e., row number in the matrix).

(5) Boxplot a suitable variable in `Mroz`.

(6) Make a new graph that boxplots that same variable, but is a split boxplot in the sense that it has several boxplots side-by-side in the same panels, with one box per each level of a factor variable. (Hint: `boxplot(split())`)

(7) Let $a = 1 : 150000$. Now write `R` commands to replace every 3rd element of a with its square. Demonstrate that the same code works on $a = 1 : 15$ by showing the input and output in `R`. In this problem and the next two, you do not need to write a **function**, but you do need to make your code sufficiently general that it would work, virtually unchanged, on very similar problems. **Also, do not use `for()` loops on these problems!**

(8) Let $a = 1 : 200$. Write commands to produce a vector b that contains only the perfect squares within `a`. Demonstrate with the input and output.

(9) Type `set.seed(1)` then `m=matrix(rnorm(15),3,5)`. Write code that rearranges m so that its rows are in increasing order of row means. (Hint: see `apply()`. Output a matrix that is this sorted version and has an extra column containing the row means.)