STATISTICS 210 – Fall 2019 – Homework 1
Handed out: Thursday September 26, 2019
Due: Thursday October 3, 2019

Reading: Sep. 26 - Oct. 3 Two-sample methods (Chap 15, App A.6-A.7)
Oct. 9 – Oct. 18 One factor ANOVA (Chap 16-18, 21)

NOTE: There is no required computation on this homework. The last item asks you to install R and RStudio in advance of a brief demonstration next Monday during our discussion section.

NOTE: Data sets (none on this HW) and handouts can be found on the course website www.ics.uci.edu/~ sternh/courses/210.

1. A health insurance company monitors the performance of their claim processing division by collecting data on the amount of time (in days) that it takes the division to process the claims. Historically the mean number of days required to process a claim is 32 days and the standard deviation is 8 days.

(a) The number of days required to process a claim only takes integer values. Do you think a normal distribution would be a suitable approximation to use for the distribution of claim processing times? Explain.

(b) Regardless of your answer in (a), use a normal distribution to estimate the proportion of claims with processing times greater than 40 days? greater than 50 days? greater than 60 days?

(c) The company installs a new software package to assist the claim processing division. After a month the manager decides to use a statistical approach to determine if the new software has improved mean processing time. Carefully describe the population of interest and the population parameter of interest.

(d) She takes a sample of 50 claims from the month and records the processing time for each of these claims. What is the distribution of the mean of these 50 claims under the hypothesis that the new software package has had no impact on the distribution of processing times? How do you know?

(e) She finds that the sample mean is 30.4. Find the probability of observing a decrease in mean processing time (relative to the historical mean) this large (or larger) under the hypothesis that the software has no impact. What conclusion do you draw about the new software?

(f) Though the mean is lower, the manager finds that 4% of the samples had processing times of 50 days or more. Does this concern you? Explain why or why not.

2. Experiment or observational study? – A study was carried out to determine the effect of receiving free milk on the achievement of children in school. The idea is that many children don’t have adequate breakfast and it might be worthwhile to give free milk. The children in one school were randomly assigned by investigators into a milk group and a control group (no milk). At the end of the year the performance of the students on a standardized reading test will be used to compare the two groups.

(a) Identify the experimental units, treatments, and response variable for this study.

(b) Is this study an experiment or observational study? Explain.

(c) Explain why it was good for investigators to randomly assign students to the milk/control groups.

(d) After the random assignment was done, some teachers switched the assignments of some children to insure that especially needy children got milk. At the end of the year, reading levels were assessed for all children.

i. Explain why the well-meaning teachers have created a data analysis problem.

ii. There are at least three ways to proceed. The ”intention to treat” approach would compare the reading levels of the students that were initially randomly assigned to the milk group to those assigned to the control group. What are the advantages and disadvantages of this approach?

iii. The ”as treated” approach would compare the reading levels of the students who ended up receiving milk to those that ended up not receiving milk. What are the advantages/disadvantages of this approach?

iv. The ”per protocol” approach would exclude data from the children who got switched and analyze the data only for those children who stayed in the group they were assigned. What are the advantages/disadvantages of this approach?
3. *p*-values – The *p*-value is a central concept for significance testing in statistics. It is often misunderstood and has recently been criticized by some as a key contributor to the lack of reproducibility of scientific studies (it appears that the results in a non-trivial fraction of scientific studies can not be reproduced).

   (a) Give a careful definition of the *p*-value. If you’d like to have a concrete situation to consider, assume we have the situation from problem (1) and are interested in testing $H_0 : \mu = 32$ vs $H_a : \mu < 32$. Suppose a *p*-value of .04 was reported for this test. Explain carefully what this *p*-value tells us.

   (b) Read ”The ASA’s Statement on *p*-Values” and the 2019 editorial in Nature on the term ”statistical significance” (both can be found on the website).

      i. What do you think - should “statistical significance” be banned?
      ii. Discuss positive and negative features of the use of *p*-values.

4. Install R and RStudio on your computer. See instructions on course website.