MPCS 53110 Foundations of Computational Data Analysis — Winter 2017

Homework 8. This problem set is due Tuesday February 28 at 5:30 pm.

Reading: Probability & Statistics with R (= P&S): **Read: chapter 7, sections 7.4–7.5 (not discussed in class)**; chapter 8, sections 8.1–8.3.

Written assignment:

- Solve the following "DO" exercises and homework problems "HW".
- "DO" exercises are strongly recommended to check your understanding of the concepts. **Do not submit them**.
- **Problems labeled "HW" must be submitted**.
- You are responsible for the material covered in **both** "DO" exercises and HW problems.

"Do" Exercises (*not* to be submitted):

1. **"DO"** Section 4.6: exercises 3 and 7, page 209.

2. **"DO"** Section 8.2: exercises 4, 6 and 10, page 291.

Homework Problems (to be submitted Tuesday February 21 at 5:30 pm):

- **Collaboration policy: If you work with others, indicate their names as part of your submission. You must answer each question by yourself without assistance. It is a violation of this policy to submit a solution that you cannot explain orally to the instructor/TAs.**
- **Looking for solutions to problems on the internet, even when acknowledged, is STRONGLY DISCOURAGED.**
- **Write out your work for every "theory" problem. If you just write your answer without showing your work, you will not receive credit.**

1. **HW** There is a laboratory with thousands of hampsters. The average weight of all the hampsters is 32 grams with a standard deviation of 4 grams. A grad student was asked by a researcher to select 25 hampsters for an experiment. However, before performing the experiment the researcher weighed the hampsters to determine if the student's selection was really a random sample or whether it was made with some bias (perhaps the hampsters selected were the most tame, which might indicate some difference among members of this group). If the sample mean of the 25 hampsters was 30.4, would this be significant evidence, at the 5 percent level of significance, against the hypothesis that the selection constituted a random sample? Show your work and explain your answer. (4 points)

2. **HW** A population distribution is known to have standard deviation 20. Determine the *p*-value of a test of hypothesis that the population mean is equal to 50, if the average of the sample of 64 observations is
    - 52.5 (1 point)
    - 55.0 (1 point)
    - 57.5 (1 point)

    Show your work.

3. **HW** The amount of alphenal varies in the capsules sold by a drug manufacturer. The manufacturer states that the mean value is 20.0 mg. A random sample of 25 pills were tested. The testing yielded a sample mean of 19.7 with a sample standard deviation of 1.3. Are the data strong enough evidence to reject the claim of the manufacturer? Use the 5 percent level of confidence. Show your work and

explain your answer. (4 points)

4. **HW** Download (WARNING: 7MB size) and read about a dataset. Answer these questions.
My boss thinks that users performing "key" authentication are more efficient, starting their day earlier, and working harder. We don't have the data necessary to test these claims, but the boss wants to reward those that login with a key. Form a hypothesis to test the claim that the mean authentication time of the users performing "key" authentication differs from the mean authentication time of the users that perform "password" authentication. Assume that the mean of the authentication times is approximately normally distributed and that both populations (key users and password users) have equal variances.
    - State the null hypothesis. (2 points)
    - Find the relevant *p*-value. (2 points)
    - Can you reject the null hypothesis? Should my boss proceed with his plan? (2 points)

5. **HW** Use the dataset from problem 4. Is there enough evidence to support the claim that the faculty, on average, tend to authenticate earlier in the day than staff? Use the mean authentication time and interpret "earlier in the day" as the lower of the two means.
    - What is your null hypothesis? (2 points)
    - What is the relevant *p*-value? (2 points)
    - What is your conclusion? (2 points)

6. **HW** This problem is an introduction to bootstrapping, a nonparametric method for simulating probability distributions from data. It is useful when there is no good parametric model for the data.
Start by entering the following commands in R:
```
> data(treering)
> D = diff(c(0, which(treering > 1.6)))
> n = length(d)
```
The dataset `"D"` represents interarrival times between certain rare events. It is natural to consider whether the rare events are produced by a Poisson process. Recall that the interarrival times in a Poisson process are independent, identically distributed exponential random variables.
    - a) Assuming that `D` is a vector of numbers independently drawn from the distribution Exp(lambda), find the MLE for lambda. (2 points)
    - b) The function `ks.test()` takes a numeric data vector, a cumulative distribution function, and any additional parameters to the CDF. Use it to test the hypothesis that `D` came from an exponential distribution with the value of lambda that you found in part (a). You should get a very low *p*-value, which is convincing evidence that `D` is not exponentially distributed. (2 points)
    - c) In the absence of a parametric model for `D`, if we can assume that the data are independently drawn, then a simple way to simulate the distribution that generated `D` is to draw from the values of `D` uniformly at random. Write a function in R that, given inputs `n` and `v`, returns a sample of length `n` from `v` drawn uniformly at random with replacement. This will be our "random number generator" for the rest of the problem. (2 points)
    - d) A simple way to test whether the values of `D` are independently drawn is to find the sample correlation coefficient between `D[1:(n-1)]` and `D[2:n]`. This is called the sample autocorrelation of `D` with lag 1.
    Let c be the sample autocorrelation of `D` with lag 1. Let C be a random variable equal to the sample autocorrelation with lag 1 of n numbers drawn independently from the random number generator you wrote in part (c). We wish to use the sample autocorrelation to test whether the values of `D` are independently drawn. Since the distribution of C may be asymmetric, the *p*-value for this test is given by $2*\min(P(c > C), P(c < C))$.
    Find the *p*-value by simulating the probabilities given in the expression. Is this evidence sufficient to reject the null hypothesis that the values of `D` are drawn independently? (6 points)

Gerry Brady, David Ramsey, Liwen Zhang, and Colin Hudler
Wednesday February 22 19:15:31 CST 2017