M/STAT 501 Fall 2022: Homework 2

Due Wednesday, Sep 14 by 5:00pm in Gradescope

Directions: For this homework, there are three different types of problems: 1) *Graded for Accuracy Problems*, 2) *Graded for Completion Problems*, and 3) *Extra Practice Problems.* Please turn in your work for the *Graded for Accuracy Problems* and the *Graded for Completion Problems*, as well as any *Sources Used* by Wednesday, August 31 by 5:00pm via Gradescope. Show work and/or provide justification for credit—neatness counts! *The Extra Practice Problems* do not need to be turned in, and the solutions will be posted for you to check and compare your work.

You are encouraged to use R Markdown for your homework, but it is not required.

Graded for Accuracy Problems (turn in)

1. The U.S. Senate consists of 100 senators—two from each of the 50 states. A Senate committee of six is to be formed.
   a. How many ways are there to form the committee from the 100 senators?
   b. If at most one senator from each state can serve on the committee, how many ways are there to form the committee?
2. A group of researchers is interested in determining the risk factors for developing breast cancer in women. Since this is an exploratory study, they plan to use a null hypothesis significance test for 15 different potential factors. However, suppose that none of the potential factors tested actually have an effect on the risk of breast cancer.[1]
   a. Suppose the researchers plan to use a significance level of $\alpha = 0.05$ for each test. Use Bonferroni's or Boole's Inequality to find an upper bound on the probability that at least one of these tests is statistically significant, i.e., at least one p-value is less than 0.05. That is, find an upper bound on the probability of at least one Type 1 error. This bound is called the "family-wise Type 1 error rate".
   b. Now assume that these 15 tests are independent. What is the probability that at least one of these tests is statistically significant if each test uses a significance level of $\alpha = 0.05$?
   c. If the researchers would like to limit the family-wise Type 1 error rate to at most 0.10, what significance level should they use for each of the 15 tests?
3. Let A and B be two events in sample space S.
   a. Prove or disprove: $P(A^C|B) = 1 - P(A|B)$.
   b. Prove or disprove: $P(A|B^C) = 1 - P(A|B)$.
   c. Prove or disprove: $If P(A|B) > P(A)$, then $P(B|A) > P(B)$.
4. Dr. Stoneguy, a wealthy diamond dealer, decides to reward her daughter by allowing her to select one of two boxes at random. Each box contains 3 stones. In one box, 2 of the stones are real diamonds, and the third stone is a worthless imitation. In the other box, one stone is a real diamond and the other 2 stones are worthless imitations. Dr. Stoneguy allows her daughter to choose one stone at random from the selected box, and to examine it to see if it is a real diamond. If the tested stone is a real diamond, the daughter will take the selected box. Otherwise, she will take the other box. What is the probability the daughter will get 2 real diamonds?

[1]For a great cartoon on the dangers of multiple testing see https://xkcd.com/882/.

Graded for Completion Problems (turn in)

Classic Hat Problem Probabilities: Suppose n people go to a fancy restaurant. Each person is wearing a hat and checks his/her hat at the door as he/she arrives. The hat-check attendant gets tipsy throughout the evening, and returns a random hat to each person as they leave. Assume the patrons leave in a random order.

1. What is the probability that no person gets his or her own hat back? What is the limit of this value as n goes to infinity?
2. What is the probability that exactly k people get their hat back? What is the limit of this value as n goes to infinity? (We will see this expression again when we look at the Poisson distribution!)

Classic Hat Problem Simulation: Now, we will use `R` to simulate this process and check the probabilities we just derived. **Goal**: Write an `R` function that will take a value of n (number of hats), then:

- Simulate the hat matching process 10,000 times.
- Keep track of the number of matches each time and store them in a vector.
- Return the vector of 10,000 "numbers of matches".

When writing a function in `R` that simulates a random process many times, use the following strategy:

- Write code to simulate the random process once for a particular value of n.
- Set up an empty vector to record output from each trial: `n.matches <- NULL`.
- Wrap the code that simulates one trial within a `for` loop, storing each trial's output in the `i`th spot of the vector:

```
for(i in 1:10000){
    [Code for a single trial here]
    n.matches[i] <- [Output from my trial]
}
```

- Wrap all of that code into a function, changing any hard-coded values into variables, e.g., use `n` rather than a particular value, and include `n` as an argument to the function:

```
my.awesome.function <- function(n, reps){
    # n = number of hats
    # reps = number of trials
    n.matches <- NULL
    for(i in 1:10000){
        [Code for a single trial here]
        n.matches[i] <- [Output from my trial]
    }
    return(n.matches)
}
```

- Now use your function to answer the following questions!

```
trials <- my.awesome.function(n = 20, reps = 10000)
```

3. Use your simulation function to estimate the probability that no one gets their hat back when you have $n = 10$ hats, $n = 20$ hats, and $n = 100$ hats. What happens to this probability as n increases?
4. Use your simulation function to estimate the probability that exactly 5 people get their hat back when $n = 20$.
5. Use your simulation function to estimate the average number of patrons who get their own hat back when you have $n = 10$ hats, $n = 20$ hats, and $n = 100$ hats. What happens to this value as n increases?

Sources Used (turn in)

At the end of your assignment, list and cite all sources used when working on this assignment, including individuals with whom you discussed the problems. If you did not use any sources besides our textbook or the course notes, please note this.

Extra Practice Problems (do not turn in, solutions posted in D2L)

1. A package of eight light bulbs contains three defective bulbs. If two bulbs are randomly selected for use, find the probability that neither one is defective.

2. Capture-recapture studies are common in ecology. One form of the study is conducted as follows. Suppose we have a population of N deer in a study area. Initially, n deer from this population are captured, marked so that they can be identified as having been captured, and returned to the population. After the deer are allowed to mix together, m deer are captured from the population and the number k of these deer having marks from the first capture is observed. Assuming that the first and second captures can be considered random selections from the population and that no deer have either entered or left the study area during the sampling period, what is the probability of observing k marked deer in the second sample of m deer?

3. Casella and Berger Exercises 1.20, 1.22, 1.36, 1.38.