Homework #3

APPM 4570/5570, Statistical Methods, Fall 2017

Due in class on Friday September 22, 2017. *Instructions for "theoretical" questions: Answer all of the following questions. The "theoretical" problems should be neatly numbered, written out, and solved. Do not turn in messy work. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.*

Theoretical Questions

1. The CU Boulder triathlon team has 11 women and 7 men. The team is going to a race and can only enter four participants. What is the probability of randomly selecting a race squad of four participants with exactly three women.

   Number of ways to choose 3 women from the group of 11:

   $$\binom{11}{3}$$

   Number of ways to choise 1 man from the group of 7:

   $$\binom{7}{1}$$

   Number of total possible combinations for selecting 4 people from a group of 18:

   $$\binom{18}{4}$$

   So, we find the probability of selecting a team with exactly 3 women to be:

   $$\frac{\binom{11}{3} * \binom{7}{1}}{\binom{18}{4}} \approx 0.378$$

2. One out of 75 people has disease D. Suppose a doctor suspects that one of her patients may have that disease. She orders a diagnostic test T for that patient. The test is not perfect—it accurately claims the disease is "present" in 90% of the patients who actually have it, and accurately declares the disease as "absent" in 80% of the patients who indeed don't have the disease.

   We are given these probabilities,

   $$\text{Probability of disease} : P(D) = \frac{1}{75} = 0.0133$$
   $$\text{Probability of positive test given disease is present} : P(T|D) = 0.90$$
   $$\text{Probability of negative test given disease isn't present} : P(\sim T| \sim D) = 0.80$$

   From these, we can deduce that,

   $$P(\sim D) = 0.98667,$$
   $$P(\sim T|D) = 0.1, \text{ and}$$
   $$P(T| \sim D) = 0.2.$$

(a) What is the probability that the test result comes back negative (the test says "absence of disease")? Does this mean your patient definitely does not have the disease?

Using the law of total probability,

$$\begin{aligned} P(\sim T) &= P(\sim T|D)P(D) + P(\sim T| \sim D)P(\sim D) \\ &= (0.1)(0.0133) + (0.8)(0.98667) \\ &\approx 0.791. \end{aligned}$$

No, it doesn't mean that the patient definitely doesn't have the D disease.

(b) If the test is negative, how likely is he to actually have disease D?

Using Bayes' theorem,

$$\begin{aligned} P(D| \sim T) &= \frac{P(\sim T|D)P(D)}{P(\sim T)} \\ &= \frac{(0.1)(0.0133)}{0.791} \\ &\approx 0.00168. \end{aligned}$$

(c) How about if the test comes back positive (the test says "disease")?

Using Bayes' theorem,

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{1 - P(\sim T)} \\ &= \frac{(0.9)(0.0133)}{1 - 0.791} \\ &\approx 0.0573 \end{aligned}$$

(d) What do you think about the accuracy of this test?

The test is not at all that reliable!

3. Suppose we are going to roll a biased 6-sided die 3 times. The die is biased so that the probability of getting 2 is twice the probability of the probability of getting a 1. The probability of getting a 3 is 3 times the probability of getting a 1. And so on up to 6 (i.e, getting a 6 is 6 times the probability of getting a 1). Define the following events:

O: "the die came up odd all three times"
E: "the die came up even exactly 1 time"
F: "the die came up greater than four all three times"
S: "the results of the three rolls add up to 17"
Calculate the following probabilities.

First, we need the probabilities of rolling a certain number (P(1), P(2), ..., P(6)). We know that,

$$p + 2p + 3p + 4p + 5p + 6p = 1 = 21p$$

where p = probability of rolling a 1. From this, $p = \frac{1}{21}$. So,

$$\begin{aligned} P(1) &= \frac{1}{21}, \\ P(2) &= \frac{2}{21}, \\ P(3) &= \frac{3}{21}, \\ P(4) &= \frac{4}{21}, \\ P(5) &= \frac{5}{21}, \text{ and} \\ P(6) &= \frac{6}{21}. \end{aligned}$$

(a) $P(E)$

First, let's find the probability that it turns up even and odd on one roll:

$$\begin{aligned} P(\text{`2 or 4 or 6'}) &= P(2) + P(4) + P(6) \\ &= \frac{2}{21} + \frac{4}{21} + \frac{6}{21} \\ &= \frac{12}{21}. \end{aligned}$$

$$\begin{aligned} P(\text{`1 or 3 or 5'}) &= P(1) + P(3) + P(5) \\ &= \frac{1}{21} + \frac{3}{21} + \frac{5}{21} \\ &= \frac{9}{21} \end{aligned}$$

Using this, we see that

$$P(E) = 3 \cdot P(\text{`even and odd and odd'}) = 3 \cdot (\frac{12}{21})(\frac{9}{21})(\frac{9}{21}) = \frac{2916}{9261} \approx 0.315.$$

(b) $P(O)$

Using the result from part (a), we see that

$$P(\text{`odd and odd and odd'}) = (\frac{9}{21})^3 = \frac{729}{9261} \approx 0.0787.$$

(c) $P(F)$

First, let's find the probability of rolling a die greater than 4:

$$\begin{aligned} P(\text{`die greater than 4'}) &= P(5) + P(6) \\ &= \frac{5}{21} + \frac{6}{21}. \\ &= \frac{11}{21} \end{aligned}$$

Using this, we see that

$$P(F) = P(\text{`die greater than 4 on all three rolls'}) = (\frac{11}{21})^3 = \frac{1331}{9261} \approx 0.144$$

(d) $P(S|E)$

Note that $S \cap E = \emptyset$. Thus, $P(S|E) = 0$.

(e) $P(S|F)$

Note that $S \subset F$ and thus, $S \cap F = S$. So,

$$P(S|F) = \frac{P(S)}{P(F)} = \frac{3(6/21)^2(5/21)}{0.1437} = 0.4058...$$

4. The game of Yahtzee is played with five fair dice. The goal is to roll certain 'hands', such as Yahtzee (all five dice showing the same number) or a full house (three of a kind and two of a kind). In the first round of a player's turn, the player rolls all five dice. Based on the outcome of that roll, the player has a second and third round, where he/she can then choose to re-roll any subset of the dice to get a desired hand.

(a) What is the probability of rolling a Yahtzee on the first round?

There are 6 ways to roll a Yahtzee: (1,1,1,1,1),...,(6,6,6,6,6). The sample space of 5 dice rolls is $(1,2,3,4,5,6)^2$, or 6^5 possible roll combinations for 5 dice. So, the probability of rolling a Yahtzee on the first round is:

$$\frac{6}{6^5} = 6^{-4} \approx 0.000772$$

(b) A small straight is defined as having 4 of the 5 dice all in a row (for example, $\{1,2,3,4,6\}$). What is the probability of rolling a small straight on the first round?

There are four possible combinations for a small straight:

(1,2,3,4) (2,3,4,5) (3,4,5,6)

Since small straights only require 4 dice, we need to account for the number of the combinations of the last die in conjunction with a small straight, making sure the 5th die does not turn the small straight into a large straight.

For the first combination, the fifth die can be any value except 5, for a total of 5 possible values. The fifth die in the second combination can be any value except 1 or 6, for 4 possible values. The fifth die in the last combination can be any value except 2, for 5 possible values. Altogether, there are 14 combinations that result in a small straight.

Now, we consider the 2 combinations where each value is unique:

$$(1,2,3,4,6) \quad (1,3,4,5,6)$$

Each of these have 5! ways to arrange the dice, for a total of:

$$2 \times 5! = 240$$

Next, the remaining 12 combinations will have a value repeated, such as (1,1,2,3,4). For these 12, we must choose 2 dice out of the 5 to take on the same value, leaving us with 3! ways to arrange the remaining 3 dice. So

$$12 \times \binom{5}{2} \times 3! = 720$$

Adding these two quantites together, we find the probability of rolling a smal straight on the first throw to be:

$$\frac{960}{6^5} = \frac{10}{81} \approx 0.123.$$

(c) (**APPM 5570 Only**) Suppose that, on the second round, the dice are $\{2, 3, 4, 6, 6\}$. You decide to re-roll both sixes in the third round. What is the probability that you roll either a small straight or a large straight (a large straight is where all five dice are in a row)?

If we decide to reroll the two 6's, then we start with the 3 dice combination of (2,3,4). For the last two dice, we can role either (5,2), (5,3),...,(5,6), (2,5), (3,5),..., (6,5), (1,1),...,(1,4),(1,6), (1,1),...,(4,1), or (6,1).
So, we can see that the probability of getting either a small or a large straight is:

$$2(1/6)(5/6) + 2(1/6)(5/6) = 5/9 \approx 0.556.$$

5. Consider the function $f(x) = \frac{1}{5}(x-5)$ on the interval $x \in [5, b]$. Find b such that $f(x)$ is a pdf.

Recognizing that a pdf must integrate to 1 over its support, we see that

$$\begin{aligned} 1 &= \int_5^b \frac{1}{5}(x-5)dx \\ &= \left[\frac{x^2}{10} - x\right]_5^b \\ &= \frac{b^2}{10} - b + 2.5. \end{aligned}$$

We solve $\frac{b^2}{10} - b + 1.5 = 0$ for b:

$$b = \frac{1 \pm \sqrt{1 - 4(1/10)(1.5)}}{2(1/10)} \approx (1.8377, 8.1622).$$

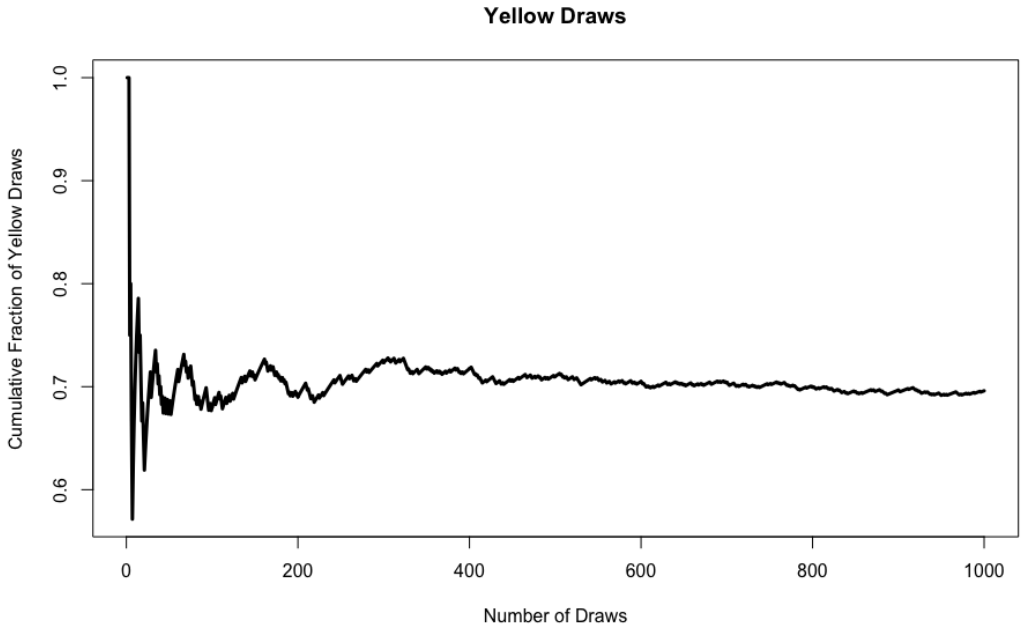
The first solution is not within the bounds; thus, $b \approx 8.1622$.

Computational Questions

Instructions for "computational" questions: Your work should be neatly done and include all graphs, code, and comments, labeled and in order based on the problem you are addressing. Do not put graphs in at the end, stick code in random locations, or do anything else that will make this homework difficult to read and grade. If you turn in something that is messy or out of order, it will be returned to you with a zero. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.

1. There is an urn that has 100 marbles in it: 2 are red, 70 are yellow, 13 are green and 15 are blue. Simulate data to imitate someone drawing a marble from the urn (replacing after each draw), with the color recorded at each draw (HINT: the `sample` function in `R` will be useful here).

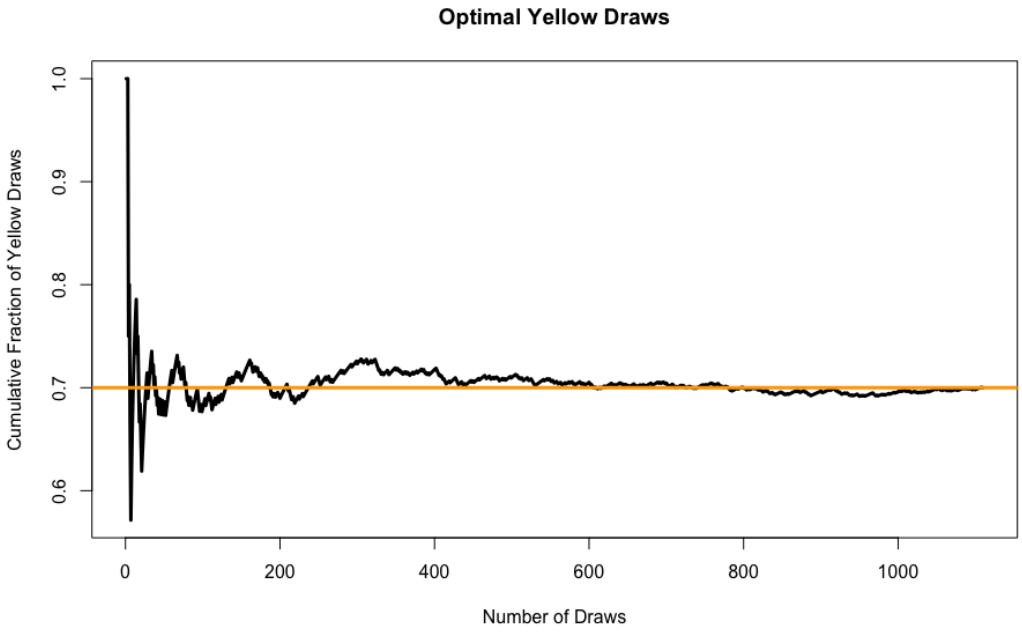
   (a) Plot the fraction of times a yellow marble is drawn in the first 1000 draws. On the x-axis, include the number of draws, and on the y-axis, include the cumulative fraction of yellow draws. Why does this number fluctuate more when the number of draws is smaller?



The figure above shows the Cumulative fractions of yellow draws for 1000 draws. The cumulative fraction fluctuates more for small number of draws since the probability of drawing a yellow is sensitive to the

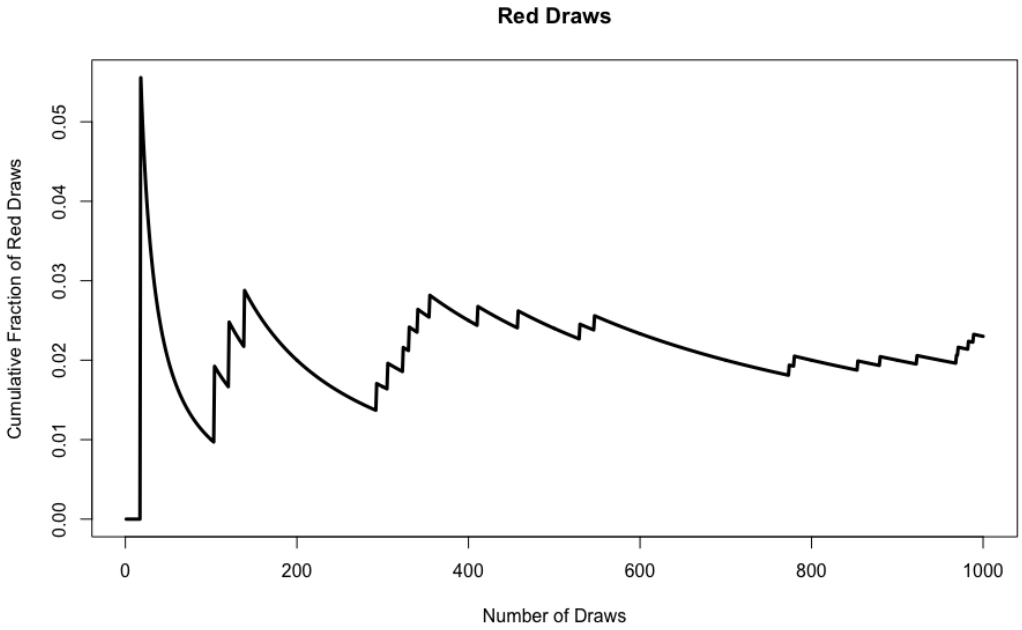
overall small number of draws. The fraction will vary largely when the sample of data is small. This behavior starts to die out as the sample size increases.

(b) How many draws are needed before the fraction of yellow marbles drawn in the simulated data is close to the probability of drawing a yellow marble?

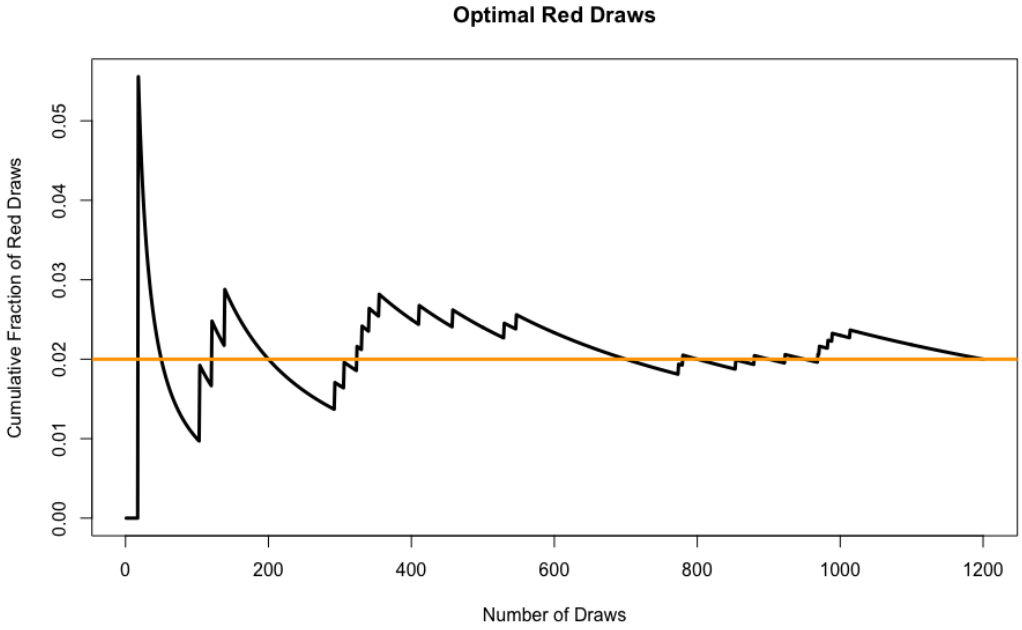


To determine the desired number of draws such that the fraction of yellow marbles drawn in the simulated data is close to the probability of drawing a yellow marble, a tolerance of 10e-6 is chosen for the difference between the true probability and numerical probability. The result is shown in the figure above with an optimal number of 1110 draws. The horizontal line represents the true probability.

(c) Repeat the two steps above for the red marble. What do you notice that is different, and why do you think it is different?



The figure above shows the Cumulative fractions of red draws for 1000 draws. Similar to the yellow draw, the cumulative fraction fluctuates more for small number of draws since the probability of drawing a red is sensitive to the overall small number of draws. Since the probability of drawing a red is small, the curve becomes smooth at some parts since no draws are occurring; however, the spikes on the figure indicate when a red was drawn.



Due to a lower probability, the desired number of draws to match the cumulative fraction of red draws to the known true probability is higher.

A larger sample size is needed to characterize a small probability. The result is shown in the figure above with an optimal number of 1200 draws. The horizontal line represents the true know probability.

2. Run a simulation that estimates the probability calculation in theoretical question 1. That is, simulate the formation of a four person race squad from 11 females and 7 males many many times, and find the relative frequency of squads that have exactly three women (HINT: again, the `sample` function in `R` will be useful here).

```
   #############
n = 10000
#function that creates one sequence of 4.
#Males are represented by zero, females by 1.
f = function(){
s = sample(c(0,1), size = 4, replace = TRUE, prob = c(7/18,11/18))
return(s)
}

#replicate the sequence many times (I like working with rows, so I transpose: t())
x = t(replicate(n, expr = f(), simplify = TRUE))

#count the number of rows in the matrix that contains exactly three women;
#divide by n.
dim((x[rowSums(x) == 3,]))[1]/n

#0.3614
```