14.310x: Data Analysis for Social Scientists
Fundamentals of Probability, Random Variables, Joint Distributions + Collecting Data

Welcome to your second homework assignment! We encourage you to get an early start, particularly if you still feel you need more experience using R. We have provided this PDF copy of the assignment so that you can print and work through the assignment offline. You can also go online directly to complete the assignment. If you choose to work on the assignment using this PDF, please go back to the online platform to submit your answers based on the output produced.

Good luck ☺!

Section 1 – Fundamentals of Probability

Unit 1 – Set Theory and Probability

1. For events A and B in S, which of the following formulas correspond to the probability that either A or B, but not both occur? (Select all that apply)

    a. $P(A)+P(B)-P(A\cap B)$
    b. $P(A)+P(B)-2*P(A\cap B)$
    c. $P(A)+P(B)$
    d. $(P(A)- P(A\cap B)) +(P(B)- P(A\cap B))$
    e. $P(A\cap B^C)+ P(A^C\cap B)$

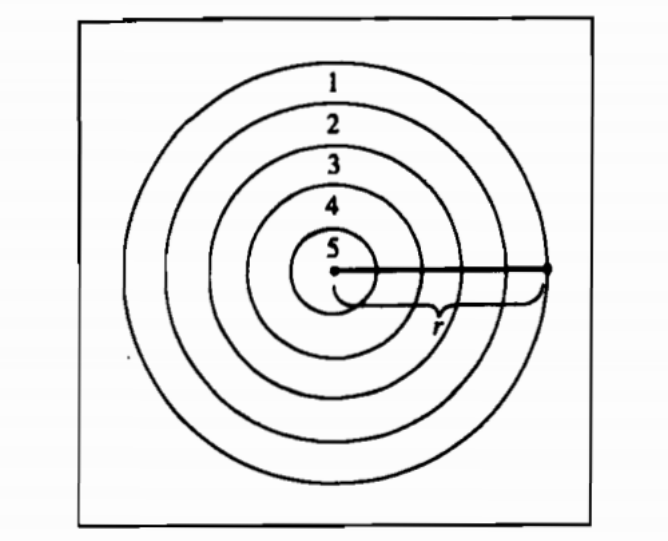
Unit 2 – Defining Probability and Examples

2. State whether the following statement is True or False: if $P(A)=1/3$ and $P(B^C)=1/4$, A and B can be disjoint.

    a. True
    b. False
    c. From the information given it is not possible to tell

3. Consider the following example taken from Casella Berger: A game of darts is played by throwing a dart at a board and receiving a score assigned to the region where the dart hits. Figure 1 shows the board and the different possible regions.

Figure 1



Assume that you are a novice player and that a friend suggests that the probability of you scoring *i* points is given by the following formula:

$$P(scoring\ i\ points) = \frac{Area\ of\ region\ i}{Area\ of\ dart\ board}$$

Does this definition satisfy the definition of probability discussed by Sara during the lecture?

a. Yes
b. No

Unit 3 – Ordered and Unordered Arrangements

4. Using an alphabet of 26 letters, how many set of initials can be formed if every person has exactly one first name and one surname (last name)?

5. In the game of dominoes, each piece is marked with two numbers. The pieces are symmetrical so that the numbered pair is not ordered: this means that (2,6) = (6,2). How many pieces can be formed with different numbers using the numbers 1, 2, …, n?

   a. 2n/(n+1)
   b. n(n+1)/2
   c. n(n-1)/2
   d. n(n+1)

Unit 4 – Independence and Bayes' Rule

Consider the example you saw in the lecture involving the Zika virus. We will start with the same set-up: A woman lives in a country where only 1 out of 1000 people has the virus. There is a test available that is positive 5% of the time when the patient does not have it, negative 1% of the time when the patient does have it, and otherwise correct. Recall that we computed that the woman's chance of having the virus, conditional on a positive test, is less than 1.9%. (By the way, in Bayesian parlance, we call the initial, unconditional, probability the "prior" and the resulting conditional probability, after updating based on observations, the "posterior.")

6. Let the conditional probability we computed (1.9%) serve the role as the new prior. Compute the new probability that she has the virus (new posterior) based on her getting a second positive test.

7. How many positive test results would she have to receive in order to be at least 95% sure that she has the virus?

   Note: You will need the correct answer from Question 6 in order to obtain the correct response for this question.

   a. Two
   b. Three
   c. Four
   d. Five
   e. No possible to infer from the available information

8. Assess whether the following statement is True or False: We obtain the same probability of having the Zika virus after a second positive test if instead of sequentially updating the conditional probability, we had used the unconditional probability and treated both tests as independent.

   a. True
   b. False
   c. No possible to infer from the available information

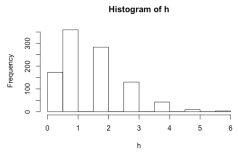
Section 2 – Random Variables, Distributions, and Joint Distribution

In R, the command to run a binomial distribution is rbinom(). Go to the R documentation and look for the arguments that are required. Then, create a sample of size 1000 using as parameters n=8, and p=0.2. Assign the created sample to the vector my_binomial. Based on your plot, answer the following questions.

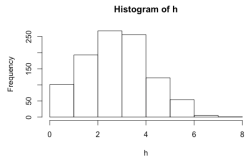
9. Write down the simplest code that will allow you to plot and histogram of the sample that you have created. Remember that the name should be my_binomial

10. Which of the following histograms is closest to the plot that you created?

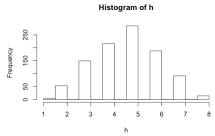
    a.



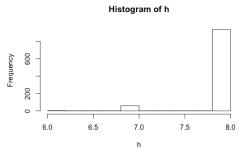
    b.



    c.



    d.



11. Now assume that instead of creating the sample with the parameter *p* equal to 0.2, this parameter is equal to one. What would be the sample mean of this new sample that you have created? And what would be the standard deviation?(Try to think about the question without coding directly in R)

12. Now, let's think about the CDF of this variable when p=0.5, and how to construct it in R. We have found a code to construct the CDF for this variable but is full of empty blanks. Can you fill them for us?

```
my_binomial <- rbinom(1000, 8, p=□)

k <- c(0:8)
cdf <- rep(0.0, times = □)
for (i in 0:8){
```

```
  j <- i + 1
  cdf[j] <- (sum(my_binomial<=□)/1000)
}

plot(k, cdf)
```

a. 0.5; 8; j
b. 0.2; 9; i
c. 0.5; 9; i
d. 0.2; 8; i
e. 0.5; 9; j

13. If a variable z follows a uniform distribution between 0 and 1, what is the value of the CDF evaluated at any value x if 0<=x<=1?

Now we will consider two independent random variables that follow a binomial distribution. X follows a binomial distribution with parameters n=8 and p=0.5; while Y follows a binomial distribution with parameters n=8 and p=0.2.

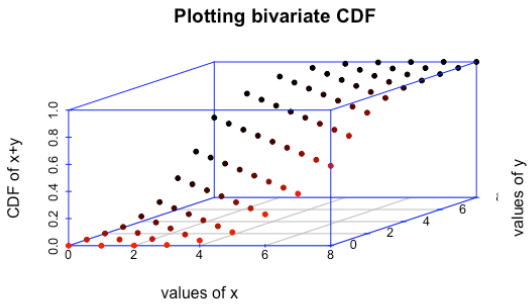
14. Taken this information into account what is the probability that the random variable X+Y has a value of zero?

One of the cool things about R is that users have developed different packages that you can use. In this example, we are going to use one of this packages called scatterplot3d. We want to plot in a three-dimensional plane the cumulative distribution of X+Y. Take a look at the code and try to understand it. Furthermore, give it a try on your computer.

```
install.packages("scatterplot3d")
library("scatterplot3d")

x <- rbinom(1000, 8, 0.5)
y <- rbinom(1000, 8, 0.5)

k <- c(rep(0:8, times=9), rep(0:8, each=9))
k <- matrix(k, ncol=2, byrow=FALSE)
z <- k[,1]+k[,2]

cdf <- rep(0.0, times = 81)

for (i in 1:81){
  cdf[i] <- sum(x+y <= z[i])/1000
}
```

```
scatterplot3d(k[, 1], k[, 2], cdf, highlight.3d = TRUE, col.axis
= "blue",
              main="Plotting bivariate CDF", pch = 20,
ylab="values of y", xlab="values of x",
              zlab = "CDF of x+y")
```

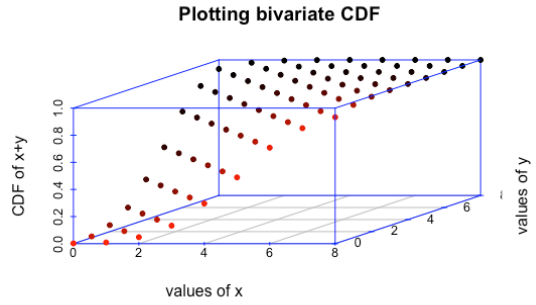
This code produces the following figure:



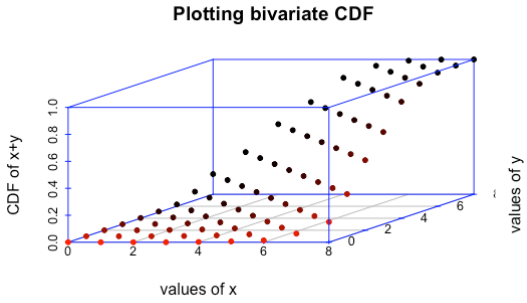
15. There is a mistake in the above code compared to what the problem states. In particular, while the code indicates Y as binomially distributed with n=8, it incorrectly sets p=0.5. Instead, you want a plot with p=0.2.

    One of your friends has given you two additional plots – one is correct with p = 0.2 and the other with p = 0.8. Your friend, however, forgets which plot is which. Take a look at the two plots below and determine which one corresponds to **p=0.2**. (You can try to solve the question analytically or use the help of R to check your answer).

    a.



    b.


Plotting bivariate CDF
CDF of x+y
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
values of x
values of y