STAT 503 – Statistical Methods for Biology

Practice Exam 1

This practice exam is longer than the real test and may take longer than 2 hours to complete.

The rules will also be different.

85 Points

Name: ____

9:00 AM 10:30 AM Online (circle one)

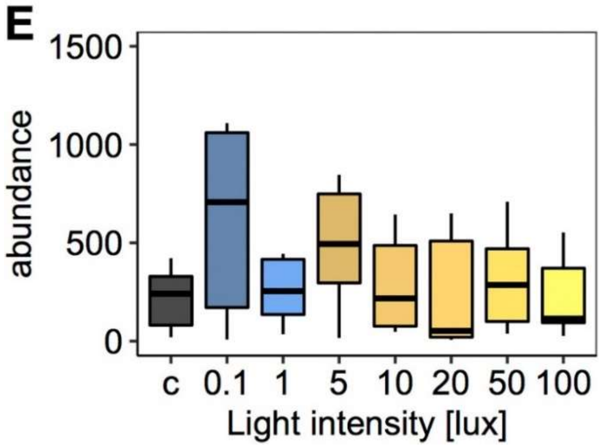
Please:

1. Do not cheat (don't speak to anyone other than me or a proctor during the exam, and don't look at anyone else's paper).
2. Write clearly and neatly. If I cannot read your answer, it will be counted wrong.
3. **Show your work** on all calculations. At a minimum, write the symbolic equations you are applying.
4. **Round final answers to 3 decimal places**.
5. Answer questions in the space provided. If you need additional space, clearly mark where the continuation may be found on both the main exam and at the location of the continuation.
6. If you have a question or need scrap paper, raise your hand and I or a TA will come to you.
7. Complete sentences are **NOT** required. Please write only as much as required to completely answer the question that was asked.
8. Phones, computers, and other electronics that are able to access the internet are **not** allowed. Any use of such a device during the exam may result in a score of zero on the test.
9. **You are allowed to bring two (2) double-sided, 8.5 × 11 sheets of notes**. Notes may be handwritten or typed in any font size, and can contain any material that you think will be helpful.
10. You may use a calculator (you may **NOT** use your phone as a calculator). Graphing calculators are allowed.
11. Please bring your Purdue Student ID to the exam and be prepared to present it when you turn the exam in.
12. If you finish early, you may turn in your exam and leave. Please be as quite as possible to minimize distractions for your classmates.
13. You will have 2 hours to complete the exam.

Question	**Page**	**Points available**	**Points earned**
1-2	3	13 (3 + 10)	
3	4	8	
	5	10	
4	6	9	
	7	8	
5	8	10	
6	9	5	
7	10	11	
8-11	11	7 (1 + 1 + 1 + 4)	
12-13	12	2 (1 each)	
14-17	13	4 (1 each)	
Total		**85**	/ 80 = __________%

1. [3 points] Complete this sentence in your own words: Two events, A and B, are independent if knowing that A occurs tells us…
   nothing about the probability that B will occur, and vice versa.

2. [10 points] The following graph shows one panel from a figure that I found in a recent issue of the journal *Current Biology*. I have edited the caption.



Cumulative abundance of the grain aphid *Sitobion avenae* feeding on barley in mesocosms without light treatments (c, control) and in different treatments with increased light intensities at night (0.1, 1, 5, 10, 20, 50, and 100 lux). Data were collected after 10 aphid generations.

   a. Please identify which variable in this study is the explanatory variable and which is the response.
      Explanatory = light intensity
      Response = aphid abundance

   b. In this graph, is light intensity considered to be a categorical variable or a numeric variable? How can you tell?
      Categorical. This is a boxplot, and boxplots always have a categorical explanatory variable. If light intensity were numeric, this would be a scatterplot. [other answers that get at light intensity being grouped or binned are also acceptable]

   c. What percentage of the data at 5 lux fall below an abundance of 500 aphids?
      Approximately 50% (the median line is ≈ 500)

   d. Are the data at 20 lux approximately symmetric, right skewed, or left skewed?
      Right skewed. The median is closer to 25% than 75%, and the lower whisker is shorter than the upper whisker.

3. [18 points] You are at a reception. Unfortunately, it's flu season, and two different strains of influenza are also attending the reception. Not including you, suppose that 12.5% of the people at the reception are infected with strain A, and 7.5% of the people are infected with strain B. The probability that a person is infected with both A and B is 0.009375.

a. Let event *A* indicate that a randomly selected attendee is infected with strain A, and event *B* indicate that a randomly selected person is infected with strain B. Are events *A* and *B* independent (show your work)?

If A and B are independent, then $P(AB) = P(A)P(B)$. All three of these quantites are given in the problem, so this can be checked:

$$P(AB) = P(A)P(B)$$
$$0.009375 = (0.125)(0.075)$$
$$0.009375 = 0.009375$$

Therefore, they are independent.

If you shake hands with a person who is infected with strain A, your probability of being infected is 0.23 (call this event *a*). If you shake hands with someone who is infected with strain B, your probability of being infected is 0.03 (event *b*). Assume that events *a* and *b* are independent.

See next page for a full tree diagram of all possible (and several impossible) outcomes for this problem. Building this tree is not necessary to solve any of the questions here. It is provided for illustration.

b. You shake one randomly selected person's hand. What is the probability that you will be infected with strain A?

The words, "If you shake hands with a person who is infected with strain A," mean that this is a conditional probability. The probability that you are infected given the person has strain A is 0.23.

$$P(Aa) = P(a|A)P(A) = 0.23 \times 0.125 = 0.029$$

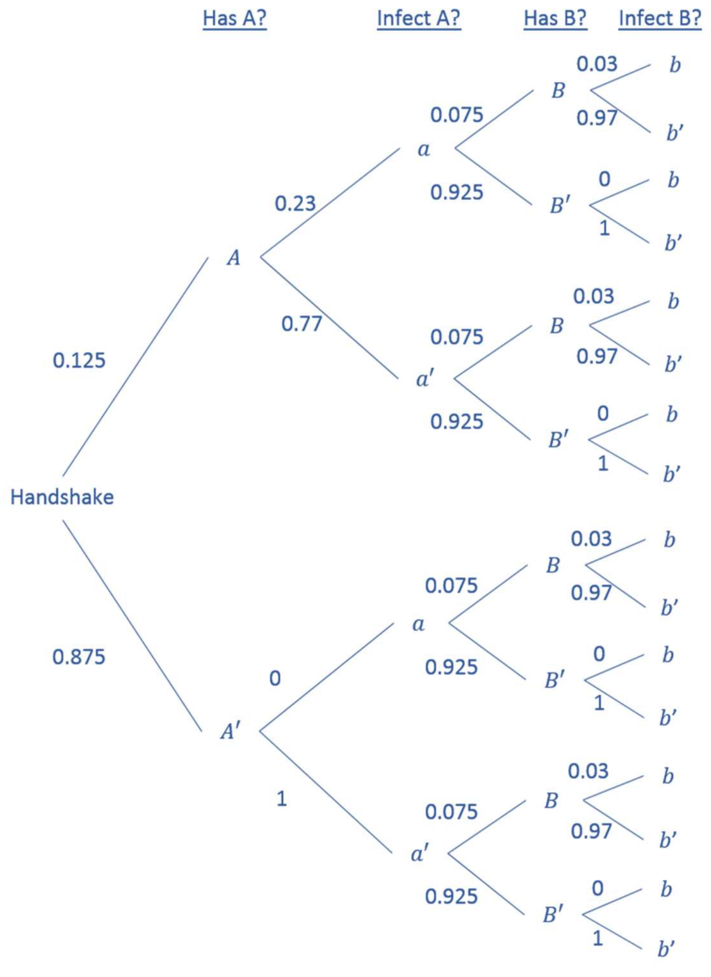
A tree for this question would look like the first two sets of forks in the tree on the next page.

c. You shake one randomly selected person's hand. What is the probability that you will **not** be infected with strain B?

The probability that you are not infected with B is $1 - P(\text{you are infected with B})$, and the probability that you are infected is the intersection of B and b.

$$1 - P(Bb) = 1 - P(b|B)P(B) = 1 - (0.03 \times 0.075) = 0.998$$

This tree would look similar to the tree for 3b., using B and b instead of A and a.



Intersection	Probability	Interpretation
$AaBb$	$P(AaBb) = 6.464 \times 10^{-5}$	Person has A and B; you are infected with A and B
$AaBb'$	$P(AaBb') = 0.002092$	Person has A and B; you are infected with A only
$AaB'b$	$P(AaB'b) = 0$	Person has A only; you are infected with A and B
$AaB'b'$	$P(AaB'b') = 0.02659$	Person has only; you are infected with A
$Aa'Bb$	$P(Aa'Bb) = 0.0002166$	Person has A and B; you are infected with B only
$Aa'Bb'$	$P(Aa'Bb') = 0.007002$	Person has A and B; you are not infected
$Aa'B'b$	$P(Aa'B'b) = 0$	Person has A only; you are infected with B
$Aa'B'b'$	$P(Aa'B'b') = 0.08903$	Person has A only; you are not infected
$A'aBb$	$P(A'aBb) = 0$	Person has B only; you are infected with both A and B
$A'aBb'$	$P(A'aBb') = 0$	Person has B only; you are infected with A only
$A'aB'b$	$P(A'aB'b) = 0$	Person has neither A nor B; you are infected with A and B
$A'aB'b'$	$P(A'aB'b') = 0$	Person has neither A nor B; you are infected with A only
$A'a'Bb$	$P(A'a'Bb) = 0.001969$	Person has B only; you are infected with B only
$A'a'Bb'$	$P(A'a'Bb') = 0.06366$	Person has B only; you are not infected
$A'a'B'b$	$P(A'a'B'b) = 0$	Person has neither A nor B; you are infected with B only
$A'a'B'b'$	$P(A'a'B'b') = 0.8094$	Person has neither A nor B; you are not infected

d. Over the course of the evening, you shake 10 people's hands. Assume they were selected randomly. What is the probability that you escape the reception without being infected by either strain?

The strains are independent, so using answers from 3b and 3c, the probability of not being infected by 1 handshake is:

$$\begin{aligned} P(not\ infected) &= P([Aa]' \cap [Bb]') \\ &= \big(1 - P(Aa)\big)\big(1 - P(Bb)\big) \\ &= (1 - 0.02875)(0.99775) \\ &= 0.969065 \end{aligned}$$

This gives:

$$P(not\ infected\ by\ 10\ people) = P(not\ infected)^{10} = 0.969065^{10} = 0.730$$

e. Alas, you were not so lucky. Assume that you were only infected by one person. Given that you come down with the flu, what is the probability that you have strain A?

This question asks you to find a conditional probability, $P(strain\ A|infected)$. Using Bayes' theorem and writing the probabilities in terms of the notation given above, this is,

$$\begin{aligned} P(Aa|[(Aa) \cup (Bb)]) &= \frac{P(Aa)}{(Aa) \cup (Bb)} \\ &= \frac{P(Aa)}{P(Aa) + P(Bb) - P\big((Aa) \cap (Bb)\big)} \\ &= \frac{P(Aa)}{P(Aa) + P(Bb) - P((Aa) \cap (Bb))} \end{aligned}$$

The numerator and first two parts here are available from 3b and 3c. Because the strains are independent,

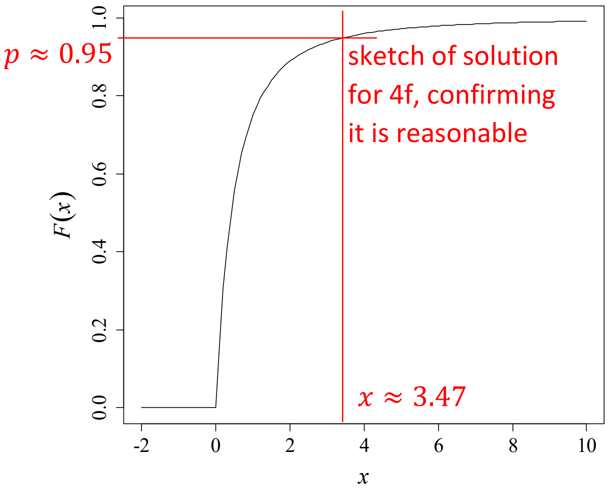
$$\big((Aa) \cap (Bb)\big) = P(Aa)P(Bb)$$

So,

$$\begin{aligned} P(Aa|[(Aa) \cup (Bb)]) &= \frac{P(Aa)}{P(Aa) + P(Bb) - P(Aa)P(Bb)} \\ &= \frac{0.02875}{0.02875 + (1 - 0.99775) - (0.02875(1 - 0.99775))} \\ &= \frac{0.02875}{0.02875 + (0.00225) - (0.02875(0.00225))} \\ &= 0.929 \end{aligned}$$

4. [14 points] An entomologist interested in pollinator behavior follows several butterflies and records their locations every 30 seconds. Using data on one of the butterflies, she finds that the distance in meters between two successive locations (referred to as *step length*) can be modeled with a random variable, X, whose cumulative distribution function follows the asymptotic curve shown to the side. The equation for this curve is:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - \dfrac{1}{(x+1)^2} & x \geq 0 \end{cases}$$



Please use the equation and curve to answer the following questions.

a. Is this a valid cumulative distribution function? What criteria are you using to make your determination?

Yes. It is (1) bounded by 0 and 1, and (2) strictly increasing, so it meets the requirement that a CDF be strictly non-decreasing.

b. What statistical population is being modeled by the random variable X in this problem? Please be as specific as you can.

The statistical population in this question is the set of all possible 30 second step lengths made by the specific butterfly whose data are modeled. Other butterflies might have different movement patterns.

c. Is the random variable X discrete or continuous? Is this an appropriate choice? Why or why not?

X is continuous (I can tell this from the graph, or from the fact that plugging any two different real numbers for x in for $F(x)$ gives me a different value; i.e., $F(x)$ is not a step function). Using a continuous random variable is appropriate because the measured variable is distance, which is continuous.

d. According to this model, what is the probability that the butterfly moves exactly 4 meters between a pair of successive locations?

For any continuous variable and any value of x, $P(X = x) = 0$. Therefore, $P(X = 4) = 0$.

e. What is the probability that the butterfly travels no more than 1 meter in a single step?

$$P(X \leq 1) = F(1) = 1 - \frac{1}{(1+1)^2} = 1 - \frac{1}{4} = 0.75$$

f. Find the value of x that represents the 95% percentile of step length.

Let $p = 0.95$, then solve for x:

$$F(x) = p$$

$$1 - \frac{1}{(1+x)^2} = 0.95$$

$$\frac{1}{(1+x)^2} = 0.05$$

$$(1+x)^2 = \frac{1}{0.05} = 20$$

$$1 + x = \sqrt{20}$$

$$x = \sqrt{20} - 1$$

$$= 3.472$$

g. [3 points] The density function for this model is $f_x(x) = r(x+1)^{-(r+1)}$, where r is a parameter that controls the average step length. **Suppose that you have two data points: $\boldsymbol{x_1 = 0.3}$ and $\boldsymbol{x_2 = 2.1}$ meters.**

(i) Write an equation for the likelihood of r, given these two data points.

$$L(r|x) = \prod f_x(x|r) = \boldsymbol{r(0.3+1)^{-(r+1)} \times r(2.1+1)^{-(r+1)}}$$

(any reasonable representation of this equation would be acceptable as an answer)

(ii) Does $r = 2$ or $r = 1.2$ provide a better fit between the model and these data? If you are short on time, explain how you would address this question.

Plugging in $r = 2$, we get: $2^2(1.3 \times 3.1)^{-3} = 4(0.015279) = 0.0611$

with $r = 1.2$: $1.2^2(1.3 \times 3.1)^{-2.2} = 1.44(0.04659) = 0.0671$

Therefore, $L(r = 1.2|x) > L(r = 2|x)$, so $r = 1.2$ provides a better fit.

Saying "calculate the likelihood for each value of r and pick the higher one" would get most of the points here. You could also use the log-likelihoods if you wanted too.

5. [10 points] The following data describe blood sodium concentrations for a number of participants in a study on diet and blood pressure, in mEq/L.

ID	[Na] (mEq/L)
1	145.3
2	141.8
3	142.0
4	142.4
5	140. 4
6	143.2

Please calculate or provide the following statistics for this dataset:

a. Sample mean

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{6}(855.1) = 142.517 \text{ mEq/L}$$

b. Sample standard deviation

$$s_x = \sqrt{s_x^2} = \left(\frac{1}{n-1}\sum (x_i - \bar{x})^2\right)^{0.5} = \left(\frac{1}{5}(13.48833)\right)^{0.5} = 1.643 \text{ mEq/L}$$

c. Sample median

The rank order of the points are (5, 2, 3, 4, 6, 1). n is even, therefore, the median is the mean of the two middle points:

$$M_x = \frac{142.0 + 142.4}{2} = 142.2 \text{ mEq/L}$$

d. Sample range

(140.4, 145.3)

6. [5 points] Suppose that the study in Question 5 is being conducted at a major teaching hospital. The participants were recruited from the hospital's nursing staff, and all participants were randomly assigned to one of two strictly controlled diets.

   a. Is this study observational or experimental? How can you tell?

      Experimental. Diet is randomly assigned.

   b. Suppose that we find an association between diet and blood pressure in the sample data. Given your answer in 6a, can we infer that a causal relationship exists between diet and blood pressure?

      Yes.

   c. Can the results of this study be used to make inferences about diet effects on blood pressure in the general public? Why, or why not?

      No. The sample is not random.

      The study participants were recruited from the hospital's nursing staff. Medical workers can be expected to be more aware of health-related issues than the general public, and may behave differently as a result. As a profession, nursing also tends to be dominated by one sex (women), may involve odd/long hours, requires a lot of time on your feet, and may be relatively stressful. All of these factors could hypothetically influence blood pressure and might also make nurses unrepresentative of the general population.

      Counterarguments are possible, and would be evaluated on their own merits.

7. [11 points] A sample of seed germination times has a mean $\bar{y} = 48.7$ days, a median of $M_y = 49.1$ days, and a standard deviation $s_y = 13.3$ days.

a. What are the median and variance of germination time in weeks?

Let $y' = y \times \frac{1\ week}{7\ days}$

$$M_{Y'} = \frac{M_y}{7} = \frac{49.1}{7} = 7.014 \text{ weeks}$$

$$s_{y'}^2 = \left(s_{y'}\right)^2 = \left(\frac{s_y}{7}\right)^2 = \left(\frac{13.3}{7}\right)^2 = 3.61 \text{ weeks}^2$$

b. My friend says that the mean of the natural log of germination time is $\ln(48.7) \approx 3.89$. Is he correct (yes or no)?

No. The natural log is a nonlinear transformation. Therefore, $\frac{1}{n}\sum \ln(y_i) \neq \ln\left(\frac{1}{n}\sum y_i\right)$.

c. Are the data in this study approximately symmetric or asymmetric, or is it impossible to tell?

They are approximately symmetric.

d. Please briefly explain the reasoning that you used to answer 7.c.

The mean and median are approximately equal ($\bar{y} \approx M_y$). In particular, their difference is much smaller than the standard deviation ($|\bar{y} - M_y| \ll s_y$), so a *z*-score for the median would be very close to zero (it is actually 0.0301).

e. Regardless of your answer in 7.c., if the data are asymmetric, in which direction are they skewed (right or left)?

Left skewed ($\bar{y} < M_y$).

f. Suppose n = 12. Please estimate a 95% confidence interval for germination time.

Standard error: $SE(\bar{y}) = \frac{s_y}{\sqrt{n}} = \frac{13.3}{\sqrt{12}} = 3.839379$ days

There are $n - 1 = 12 - 1 = 11$ degrees of freedom.

For t^*, we want the $\left(1 - \frac{\alpha}{2}\right) \times 100\%$ quantile $\rightarrow 1 - \frac{0.05}{2} = 0.975$

95% CI:

$$\begin{aligned} CI_{0.95}(\mu) &= \bar{y} \pm t^*_{n-1,0.975}\frac{s_y}{\sqrt{n}} \\ &= 48.7 \pm 2.20\ (3.839379) \\ &= (40.253, 57.147) \text{ days} \end{aligned}$$

(the answer is slightly different with a more precise value of t^*)

8. [1 point] If you want a measure of location that is not strongly affected by outliers, what sample statistic should you use?

   the median

9. [1 point] Genotype is a **categorical variable**. In R, what data type should it be represented by?

   a factor

10. [1 point] In R, the following statement will produce an error: `2x <- 37 * 2`. What is the problem?

    variable names cannot start with a number

11. [4 points] I have a data frame called mydata, which has columns named x and y. Please briefly explain the results of the following commands (what will they do)?

    a. arrange(mydata, x)

       sort the rows of mydata in increasing order of the variable x

    b. mutate(mydata, z = (x > 3))

       add a variable (or column) named z that is TRUE for rows where x is > 3 and FALSE otherwise

    c. filter(mydata, x > 3)

       return a new data frame that only contains the rows in mydata for which x > 3 is true

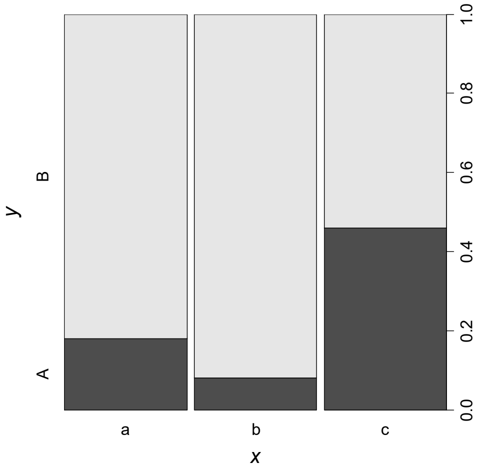
    d. select(mydaya, x)

       return a new data frame that only contains the column from mydata named x

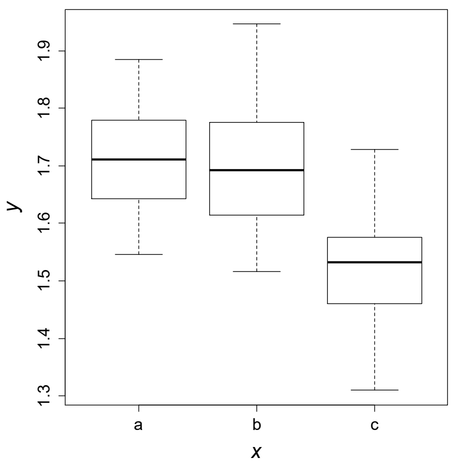
12. [1 point] Which of the following graphs illustrates a potential **association between two categorical variables**?

Plot A. All of the others have continuous (or at least numeric) response variables.

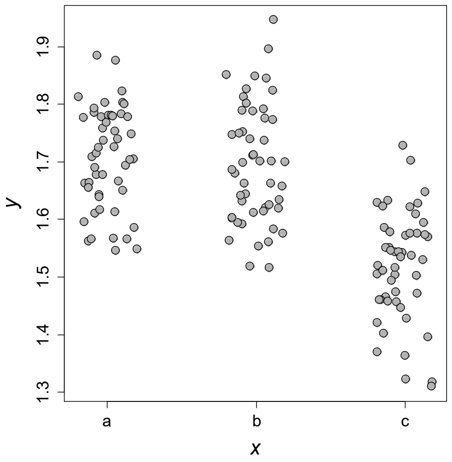
A.



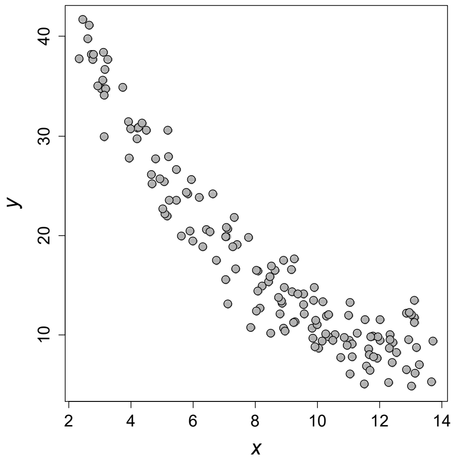
B.



C.



D.



13. [1 point] What is the **name of the data frame** that holds the data in the following line of R code?

```
xyz %>% group_by(z) %>% mutate(a = log(x))
```

xyz

14. [1 point] A friend comes to you for help with R. They are trying to make a plot with the ggplot() function, and they have installed the *ggplot2* package on their computer, but when they try to make the plot, they get the following response in the console:

```
> ggplot(mydata, aes(x = nuecleotide, y = abundance))
Error in ggplot(mydata, aes(x = nuecleotide, y = abundance)) :
  could not find function "ggplot"
```

What is the **first potential solution** that you should check?

Check to see if they ran library(ggplot2)

15. [1 point] In the following line of code, **what does frq_hz refer to**?

mean(mydata$frq_hz)

frq_hz is a column in the data frame named mydata

16. [1 point] Which of the following probability statements is **true**?

A. $P(A) = P(A')$ if A and A' are independent

B. $P(A|B) = P(A) + P(B)$

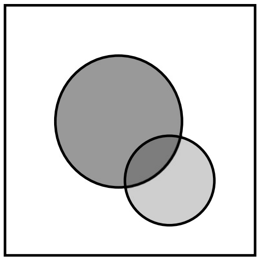
[✓ C. $P(B \cap A) = P(B|A)P(A)$]

D. $P(A) = P(A|B)$ if A and B are disjoint

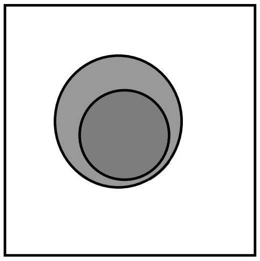
E. None of the above are true.

17. [1 point] Which of these Venn diagrams illustrates a pair of **disjoint** (mutually exclusive) events?

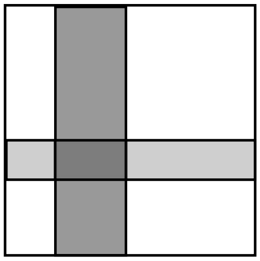
A.



B.



C.



[✓ D.]