

# Study Guide #1 – Spring 2018

STAT 110: Exam #1

Name: SOLUTION

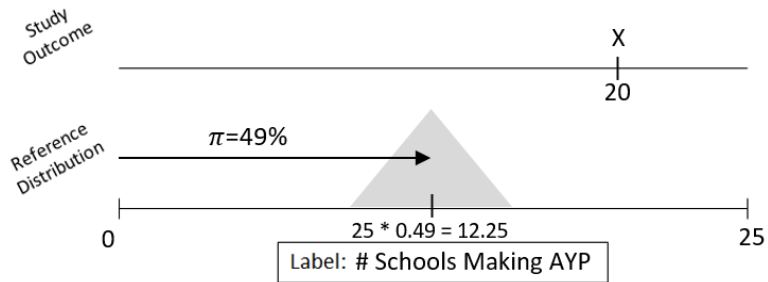
Consider the following investigation regarding South Washington County schools. The goal is to determine whether or not South Washington County should be labeled as a high performing school district.

	A	B	C	D
1	School District	Making AYP	Not Making AYP	Total
2	ST. PAUL PUBLIC SCHOOL DISTRICT	20	70	90
3	MINNEAPOLIS PUBLIC SCHOOL DIST.	14	73	87
4	INTERMEDIATE SCHOOL DISTRICT 287	34	28	62
5	ANOKA-HENNEPIN PUBLIC SCHOOL DIST.	25	29	54
6	ROCHESTER PUBLIC SCHOOL DISTRICT	15	20	35
7	ROSEMOUNT-APPLE VALLEY-EAGAN	11	23	34
8	DULUTH PUBLIC SCHOOL DISTRICT	13	14	27
9	OSSEO PUBLIC SCHOOL DISTRICT	7	20	27
10	ELK RIVER PUBLIC SCHOOL DISTRICT	17	8	25
11	SOUTH WASHINGTON COUNTY SCHOOL DIST	20	5	25
12	BURNSVILLE PUBLIC SCHOOL DISTRICT	5	16	21
13	MOUNDS VIEW PUBLIC SCHOOL DISTRICT	8	12	20
14	Entire State of MN	1283	1341	2624

Question of Interest: Is there statistical evidence to say South Washington County is performing better, i.e. more schools Making AYP, than Minnesota’s overall rate of 49%? (2 pts each)

1. Identify the smallest possible value, largest possible value, location of the pyramid, and the outcome from the study for this situation on the number lines below. (5 pts)

- Smallest possible value
- Largest possible value
- Location of pyramid
- Outcome from study



2. Next, provide details for the setup of a simulation that would allow us obtain a total of 250 simulated outcomes to investigate the research question stated above. (3 pts)

Probability of success ( $\pi$ ):

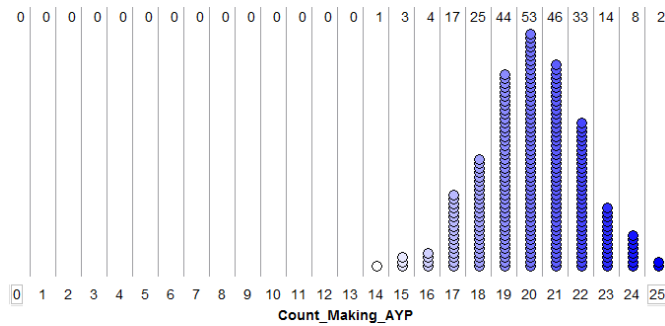
Sample size ( $n$ ):

Number of samples:

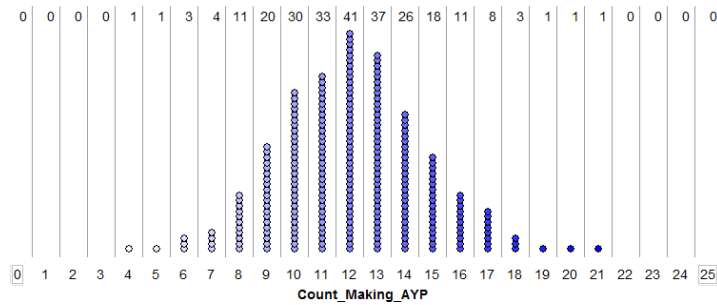
## Study Guide #1 – Spring 2018

3. The following graphs were obtained from a simulation. A total of 250 repeated trials were obtained for each graph. Which graph is most appropriate graph for the above question of interest regarding South Washington County? (3 pts)

a.



**b.**



4. The p-value is defined to be the proportion of outcomes that are as extreme as or even more extreme than our observed outcome that would provide evidence to support the research question. For South Washington County, we had 20 schools that were found to be Making AYP status. Use the graph you selected above to compute the appropriate p-value. (3 pts)

The appropriate p-value is  $\frac{\# \text{ dots that are at 20 or above}}{\text{Total \# dots}} = \frac{2}{250} = 0.008$ .

5. Which of the following is the most correct statement regarding our research question. (3 pts)
- The p-value is less than 0.05; thus, there is enough statistical evidence to say that South Washington is performing better than the state average of 49%.
  - The percentage of schools Making AYP for South Washington County is  $20/25 = 80\%$  which is greater than the state average of 49% which implies there is enough statistical evidence to say that South Washington County is performing better than the state average of 49%.
  - The percentage of schools Making AYP for South Washington County is  $20/25 = 80\%$  which is larger than 0.05; thus, we do not have enough statistical evidence to say that South Washington County is performing better than the state average of 49%.
  - Both a and b. [not b because saying the study outcome is above the expected is not enough, the study outcome must be an outlier – not just above]

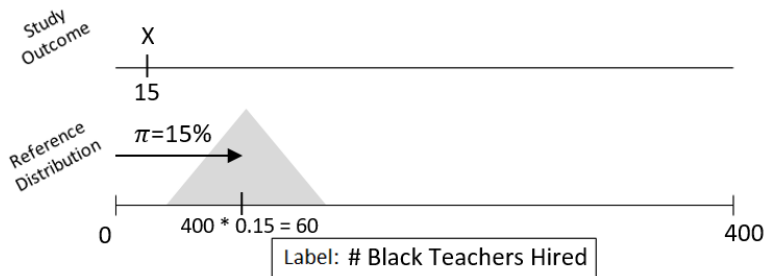
## Study Guide #1 – Spring 2018

In 1977, the U.S. government sued the City of Hazelwood (a suburb of St. Louis) on the grounds that the district had discriminated against black persons in its hiring of school teachers. The statistical evidence introduced noted that of the 400 teachers hired in 1972 and 1973 (the years following the passage of the Civil Rights Act), only 15 were black. According to 1970 census figures, however, about 15% of teachers employed in the surrounding areas in St. Louis County that year were black.

Research Question: Is there statistical evidence to suggest that the City of Hazelwood discriminated against blacks in the hiring of school teachers during this time period?

6. Identify the smallest possible value, largest possible value, location of the pyramid, and the outcome from the study for this situation on the number lines below. (5 pts)

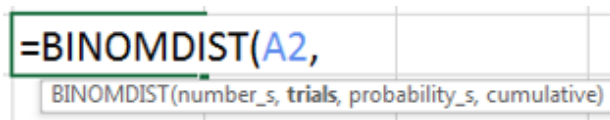
- Smallest possible value
- Largest possible value
- Location of pyramid
- Outcome from study



Next, consider the method used to obtain the p-value in Google Sheets.

	A	B	C	D
	# Black	Binominal		
1	Teachers Hired	Probabilities	A	B
2	0	=BINOMDIST(A2,		, FALSE)
3	1	4.13335E-27		
4	2	1.45518E-25		
5	3	3.40684E-24		
6	4	5.96698E-23		
7	5	8.33971E-22		

7. The =BINOMDIST() function is used in Excel to obtain the probabilities for the reference distribution.



- a. Consider the second argument in the BINOMDIST() function, i.e. the value for Box A. Excel's terminology for the second argument is **trials**. What value should be used for **trials** in order to obtain the correct probability model for the reference distribution? (3 pts)

Value for trials: 400

## Study Guide #1 – Spring 2018

- b. Next, consider the third argument in the =BINOMDIST() function, i.e. the value for Box B. Excel's terminology for the third argument is **probability\_s**. What value should be used for **probability\_s** in order to obtain the correct probability model for the reference distribution? (3 pts)

Value for probability\_s: 0.15

8. The probability values in Excel will continue to increase as the # of Black Teachers Hired increases, but eventually the probability values will start to decrease. What is the value for # of Black Teachers Hired that will have the highest probability? Briefly discuss. (3 pts)

The peak of the pyramid will occur at 60, i.e. the expected value.  
The expected value is computed as  $400 * 0.15 = 60$ .

9. Which of the following best describes what you would do next to find the p-value in Excel? Circle your answer. (3 pts)
- a. Add up the probability values for 15 or more black teachers hired.
  - b. Add up the probability values for 15 or fewer black teachers hired.
  - c. Get the probability for 15 black teachers hired.
  - d. Add up the probability values for 60 or more black teachers hired.
  - e. Add up the probability value for 60 or fewer black teachers hired
  - f. Get the probability for 60 black teachers hired.

Questions 10-13 refer to the following scenario. With movie-viewing-at-home made so convenient by services such as Netflix, video on demand, etc., do a majority of city residents now prefer watching movies at home rather than going to the theater? To investigate, a local high school student, Lori, decides to conduct a poll of adult residents in her city. She selects a random sample of 100 adult residents from the city and gives each participant the choice between watching a movie at home or the same movie at the theater. She records how many choose to watch the movie at home.

10. Which of the following most closely represents the setup of a spinner in Tinkerplots? (3 pts)
- a. Adult residents of the city are equally likely to choose to watch the movie at home as to watch it at the theater.
  - b. Adult residents of the city are more likely to choose to watch the movie at home than to watch it at the theater.
  - c. Adult residents of the city are less likely to choose to watch the movie at home than to watch it at the theater.
11. Which of the following most closely represents the research question? (3 pts)
- a. Adult residents of the city are equally likely to choose to watch the movie at home as to watch it at the theater.
  - b. Adult residents of the city are more likely to choose to watch the movie at home than to watch it at the theater.
  - c. Adult residents of the city are less likely to choose to watch the movie at home than to watch it at the theater.

## Study Guide #1 – Spring 2018

12. After analyzing her data, Lori finds that significantly more than half of the sample (p-value 0.012) preferred to watch the movie at home. Below are two possible explanations for the observed majority. Based on her analysis, which explanation do you find the most plausible (i.e., reasonable or believable)? (3 pts)
- a. More than half of the adult residents in her city prefer to watch the movie at home.
  - b. There is *no* overall preference for movie-watching-at-home in her city, but by pure chance her sample just happened to have an unusually high number of people choose to watch the movie at home.
  - c. (a) and (b) are equally plausible explanations.
13. Which of the following is the most valid interpretation of Lori's p-value of 0.012? (3 pts)
- a. A sample proportion as large as or larger than hers would rarely occur.
  - b. A sample proportion as large as or larger than hers would rarely occur if the study had been conducted properly.
  - c. A sample proportion as large as or larger than hers would rarely occur if 50% of adults in the population prefer to watch the movie at home.
  - d. A sample proportion as large as or larger than hers would rarely occur if more than 50% of adults in the population prefer to watch the movie at home.
14. A friend claims that he can taste the difference between Coke and Coke Zero. You set up an experiment during which your friend tastes a sample of each and identifies which sample is the Coke Zero, and this process is repeated a total of eight times. Your friend makes the correct distinction between Coke and Coke in 7 of the 8 attempts. What value for the parameter would you use when setting up a spinner in Tinkerplots to investigate whether or not your friend has the ability to tell the difference between Coke and Coke Zero? (3 pts)
- a.  $\pi = 1/2$
  - b.  $\pi = 1/8$
  - c.  $\pi = 7/8$
  - d. It is impossible to tell.
15. Consider the previous problem. You use the binomial distribution and find that the probability he could get 7 or more correct by guessing is 0.035. Which of the following is the most correct? (3 pts)
- a. We do not have enough statistical evidence to say that he can really tell a difference (p-value = 0.035).
  - b. We do have enough statistical evidence that he can tell a difference because he got more than 4 correct which is the expected number if he was guessing. (p-value = 0.035).
  - c. We do have enough statistical evidence to say that he can tell a difference (p-value = 0.035).
  - d. Both b and c.
16. Consider the previous problem. Which of the following is the most correct interpretation of the p-value, which was .035 or 3.5%? (3 pts)
- a. There is a 3.5% chance that your friend is guessing.
  - b. There is a 3.5% chance that your friend can tell the difference between Coke and Coke Zero.
  - c. If your friend is just guessing, there is a 3.5% chance he would get 7 or more correct.
  - d. If your friend can tell the difference between Coke and Coke Zero, there is a 3.5% chance he would get 7 or more correct.

## Study Guide #1 – Spring 2018

17. Recall the Swain vs. Alabama example. The expected number of blacks in the juror pool was 26, but Swain only had 8. The p-value for testing whether the selection process of jury panel members was biased against black persons was found to be 0.0000047. The statements below present three different interpretations of this  $p$ -value. Indicate whether or not each of the following interpretations is valid or invalid. (2 pts each)

The p-value of 0.0000047 is the probability of selecting 8 (or fewer) black persons for the jury panel if the selection process is <u>truly unbiased</u> .	VALID <input checked="" type="radio"/> INVALID <input type="radio"/>
The p-value of 0.0000047 is the probability of selecting 8 (or fewer) black persons for the jury panel if the selection process is truly discriminatory.	VALID <input type="radio"/> INVALID <input checked="" type="radio"/>
The p-value of 0.0000047 is the probability that the selection process is truly unbiased.	VALID <input type="radio"/> INVALID <input checked="" type="radio"/>

**NOTE:** For those in my MWF class, I misspoke in class today when giving the answer to Problem #17. The first is valid as this is the situation in which the pyramid was constructed under unbiased.

Questions 18-20 refer to the following scenario. A recent study was conducted to explore how people's dating preferences are affected by a prospective date's tendency to reveal (vs. hide) personal information. Each subject in the study was involved in an online dating service. They viewed two questionnaires that had supposedly been completed by two prospective dates: the *revealer* vs. the *hider*.

Revealer's Questionnaire						Hider's Questionnaire					
	Never	Once	Sometimes	Frequently	Choose not to answer		Never	Once	Sometimes	Frequently	Choose not to answer
Have you cheated on a tax return?				X		Have you cheated on a tax return?				X	
Have you made a false insurance claim?				X		Have you made a false insurance claim?				X	
Have you stolen anything worth more than \$100?				X		Have you stolen anything worth more than \$100?					X
Have you fantasized about hurting someone?				X		Have you fantasized about hurting someone?				X	
Have you hid an STD from a dating partner?				X		Have you hid an STD from a dating partner?					X

Note that the revealer answered all questions, while the hider selected "Choose not to answer" for two of the questions. The researchers hypothesized that dating prospects would be more likely to choose the revealer than the hider, in spite of the fact that the revealer admitted to more negative behavior (the theory is that potential dates are more afraid of those who hide information). Of the 44 subjects tested, 28 preferred to date the revealer.

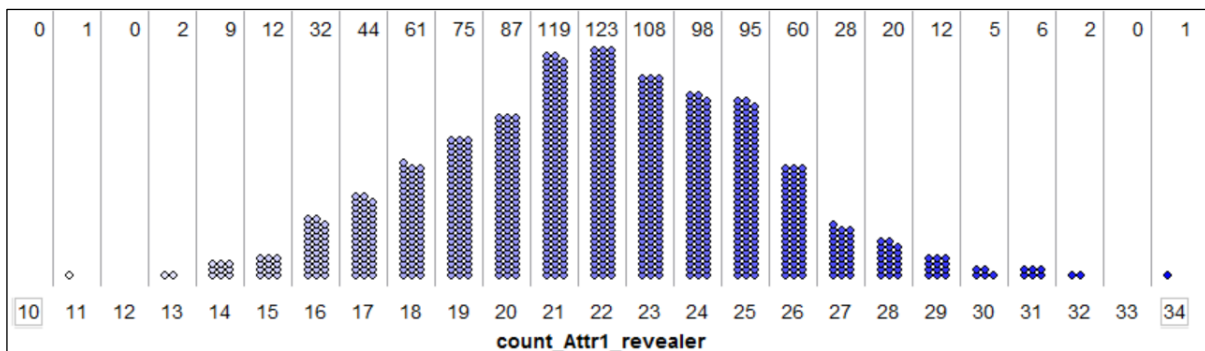
Research Question: Are potential dates more likely to choose the revealer than the hider?

## Study Guide #1 – Spring 2018

18. You have been asked to conduct a simulation study to determine what outcomes are likely to occur assuming that there is no real effect of revealing vs. hiding information. Specify the setup for the appropriate reference distribution. Suppose 1000 repeated outcomes is desired. (4 pts)

Probability of heads:   
 Number of tosses:   
 Number of repetitions:

Suppose the results of 1,000 simulated trials are as shown in the following dotplot. Note that we kept track of the number that chose the revealer in each simulated trial.



19. Recall that in the actual study, 28 subjects chose to date the revealer. Use the dotplot from the previous question to estimate the p-value for this study. (3 pts)

P-value:  $\frac{\# \text{ dots that are at 28 or above}}{\text{Total \# dots}} = \frac{46}{1000} = 0.046$ , which is 4.6%

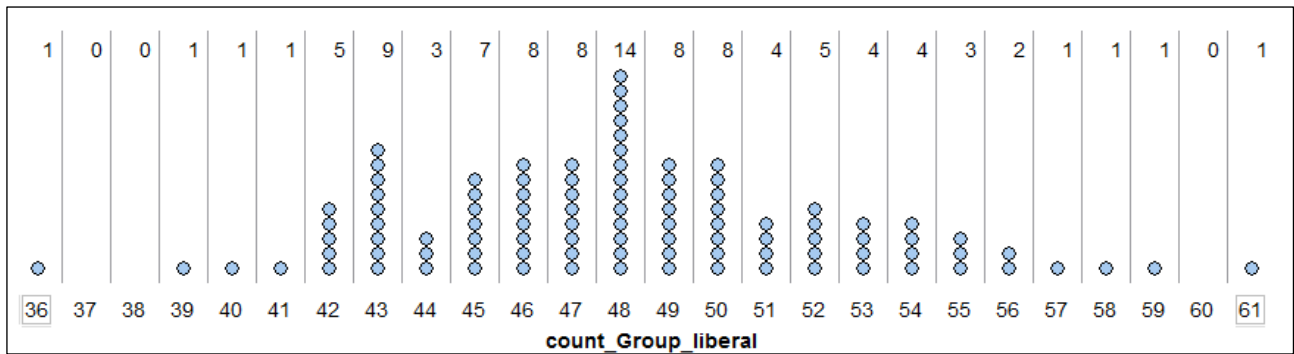
20. Which of the following is the most correct statement regarding our research question? (3 pts)
- In the study,  $28/44 = 64\%$  of the subjects chose the revealer; thus, we have statistical evidence that potential dates are more likely to choose the revealer simply because more than half of the subjects in our study did so.
  - An outcome as extreme as 28 subjects choosing the revealer happened less than 5% of the time in our simulation study; thus, there is statistical evidence that potential dates prefer the revealer.
  - An outcome as extreme as 28 subjects choosing the revealer happened more than 5% of the time in our simulation study; thus, there is statistical evidence that potential dates prefer the revealer.

## Study Guide #1 – Spring 2018

A study was conducted to determine whether Winona State students were more likely to identify themselves as liberal versus conservative. Suppose that 97 students were randomly selected from the WSU student population and were surveyed. Of the 97 students surveyed, 55 identified themselves as liberal.

Research Question: Are WSU students more likely to identify themselves as liberal than conservative?

The dotplot below shows the results of 100 trials of a simulation carried out in Tinkerplots under the assumption that exactly half of the students at WSU identify themselves as liberal (and the other half as conservative). In each of the simulated trials, the number of students that identified themselves as liberal was recorded on the plot.



21. Answer the following questions. (2 pts each)

The <del>Repeat value for this spinner</del> (For your class, the Number of tosses on pyramid app) was set to 8500 (the number of students at WSU). [Set to 97]	TRUE <input checked="" type="radio"/> FALSE
The <del>spinner</del> (For your class, the pyramid app) for this situation would represent the notion that WSU students are more likely to identify themselves as liberal than conservative. [Pyramid is 50/50, i.e. equally likely]	TRUE <input checked="" type="radio"/> FALSE
Each dot on this plot represents <i>a student</i> that identified themselves as liberal. [Dot represents the count from 97 people]	TRUE <input checked="" type="radio"/> FALSE
This is a two-tailed test because students classified themselves as either liberal or conservative. [Interest in more likely, one-tailed right]	TRUE <input checked="" type="radio"/> FALSE
Based on the 5% rule, in order to obtain statistical evidence that more WSU students identify themselves as liberal than conservative, we must see 57 or more of the students surveyed identify themselves as liberal. [Yes, 57 is the cutoff of top 5%]	<input checked="" type="radio"/> TRUE FALSE
If 48 of the students surveyed identify themselves as liberal, then we have statistical evidence that more WSU students identify themselves as liberal than conservative because this value occurred most often in our simulation.	TRUE <input checked="" type="radio"/> FALSE

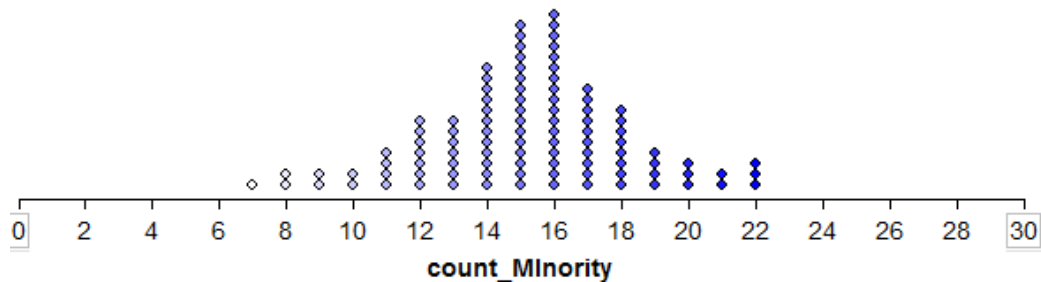


## Study Guide #1 – Spring 2018

Of interest here is whether or not minority coaches were given a fair chance to coach 3<sup>rd</sup> base in major league baseball. It was determined that 45% of the coaches in the major leagues were minority and thus if everything is fair about 45% of 3<sup>rd</sup> base coaches should be minority. The NY Times reported that only 7 of the 30 3<sup>rd</sup> base coaches are minorities and thus a potential race bias may be occurring.

Research Question: Do the data presented by the NY Times provide statistical evidence to suggest that minority coaches are not given a fair opportunity to coach 3<sup>rd</sup> base?

Consider the following outcomes from a simulation of 100 trials.



22. Which of the following is most correct regarding the above reference distribution? (3 pts)

- a. This reference distribution was constructed under the assumption that minority coaches are given a fair opportunity to coach 3<sup>rd</sup> base.
- b. This reference distribution makes no assumption about the fairness in which minority coaches are given to coach 3<sup>rd</sup> base. Making an assumption like that would be unfair and bias the analysis.
- c. This reference distribution was constructed under the assumption that minority coaches are not given a fair opportunity to coach 3<sup>rd</sup> base.

23. Answer the following True/False Questions (2 pts each)

**Definition**  
**P-Value:** the probability of observing an outcome as extreme or more extreme than the observed outcome

The context of “as extreme or more extreme” in this problem implies values of 7 or less.	TRUE FALSE
7, the observed number of minority 3rd base coaches, is the expected or anticipated value in our simulation	TRUE FALSE
The p-value for this problem is 1/100. The p-value is the proportion of dots at or below 7.	TRUE FALSE
A very small p-value, i.e. much less than 0.05, suggests the data supports the research question.	TRUE FALSE