Stat 218 - Exam 2 (Practice - KEY)

Formula sheet

	Single Variable	
	Proportion	Quantitative
Statistic	$\hat{p} = \frac{\text{Num. successes}}{n}$	$\overline{x} = \frac{\sum x}{n}$, where $\sum x$ is the sum of all observations in the sample
Standard Error	$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE(\overline{x}) = \frac{s}{\sqrt{n}}$
Standardized Statistic	$z = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$	$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$
Theory-Based CI	statistic $\pm 2 \times SE$(statistic)	

Difference in proportions	
Overall proportion of successes $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$	Standardized Statistic $z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
Theory Based CI $\hat{p}_1 - \hat{p}_2 \pm 2 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	

Concepts _____/15

Circle T (true) or F (false) for each of the following statements[1pt each]:

T If a statistic A has a smaller standard error than statistic B, the width of the 95% interval for A will be smaller than the 95% interval for B

F $2 \times$ SE is called the margin of estimation

F The midpoint of a confidence interval is the parameter value

F A confidence interval is the set of all possible values for the parameter

T Using a higher confidence level produces a wider interval of plausible values

T A representative sample is required to make unbiased inference about the population

F Sampling bias occurs when there is not random assignment

F An experiment is a study in which subjects are not randomly assigned to treatment groups or conditions

F A quasi-experiment is one that manipulates the explanatory variable and uses random assignment of treatments to groups

F A confounding variable is a variable that is related to only the explanatory variable

F The relative risk is one way to summarize the association between two quantitative variables

F We examine the difference in proportions across treatment groups because we want to know what proportion of people in the population respond to treatment.

T In order to use theory-based inference when estimating the difference in two proportions, we must have a total of 40 observations with at least 10 observations in each group

F The usual null hypothesis when comparing two proportions is $H_0 : \mu_1 - \mu_2 = 0$

F There are validity conditions for using simulation-based inference for the difference between two proportions

Plague, Inc.

You are playing a game of Plague, Inc, where you are attempting to mutate your bacterium, 'BubonicEbola' so that it can kill everyone on earth. You decide to use statistics to investigate how the game mechanics work so that you can determine the best strategy.

Infecting Madagascar _____/19

Early in the game, the goal is to evolve enough mutations that increase the infectivity of the virus, but are unlikely to be noticed (symptoms aren't too visibly obvious). If the disease becomes noticed, countries will start working on a cure and will implement quarantine measures that make it difficult for the disease to spread across the globe. The country which is seen as hardest to infect is Madagascar, because it is only accessible by boat (no airports) and will close its ports as soon as an even moderately serious disease is identified.

You play 30 games using the same starting parameters and minor mutations (coughing, sneezing, cysts) that increase infectivity but are unlikely to be noticed, and in each game, you record whether or not Madigascar is infected at the end of 3 in-game years. In total, you manage to infect Madagascar 12 times.

1. [1pt] What is the observational unit?

   A single game of Plague, Inc.

2. [1pt each] For each of the following, fill in the value. If a quantity does not apply to this problem, write NA. If a quantity is unknown, write 'unknown'. Show your work for any calculations.

   $\mu =$ *NA* $\quad \pi =$ *unknown* $\quad \overline{x} =$ *NA*

   $\hat{p} =$ *12/30 = 0.4* $\quad n =$ *30* $\quad$ successes $=$ *12*

3. [1pt] What type of variable is this?

   Categorical (or Binary)

4. [2pt] What are the validity conditions for theory-based inference on this type of data?

   At least 10 successes and 10 failures

5. [1pt] Are these conditions met? Explain.

   Yes, the validity conditions are met. We have 12 successes and 18 failures; both quantities are greater than 10.

6. [1pt] What formula should you use to calculate the standard deviation of the sample statistic?

$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

7. [1pt] What is the value of the standard deviation of the sample statistic?

$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.4(0.6)}{30}} = \sqrt{\frac{0.24}{30}} = \sqrt{0.008} = 0.0894$

8. [1pt] What formula should you use to calculate an approximately 95% confidence interval for the long-run proportion of times you can infect Madagascar using your current strategy?

$statistic \pm 2 \times SE(statistic)$

9. [1pt] Calculate an approximately 95% confidence interval for the long-run proportion of times you can infect Madagascar using your current strategy. Show your work. If you do not have a value for a component of the formula, define a variable for that value and show as much work as you can.

$\hat{p} \pm 2 \times SE(\hat{p}) = 0.4 \pm 2 \times (0.0894) = 0.4 \pm 0.1789 = (0.2211, 0.5789)$

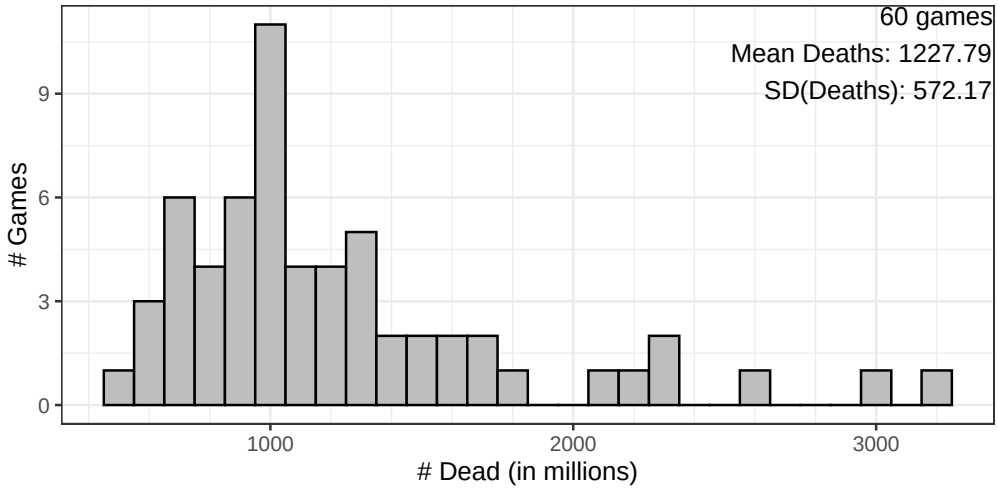
10. [2pt] Interpret your interval from 9 in the context of the problem. If you did not get an interval from 7, use (a, b) as your values and interpret as normal.
I am 95% confident that, using my strategy, Madagascar will be infected with BubonicEbola between 22.11% and 57.89% of the time over the long run.

11. [2pt] If you wanted to decrease the width of the interval you calculated in 9, while still having 95% confidence, what approach would you use? Explain why your solution will decrease the width of the interval.
Collect more data. Increasing n will decrease the width of the confidence interval by decreasing the standard error.

Number of Deaths

_____/21

Once your disease has spread across the globe, the next part of your strategy is to significantly increase the lethality of BubonicEbola by evolving symptoms such as Dysentery, Hemorrhagic Shock, and Total Organ Failure. This time, you play 60 games, using this strategy starting at the beginning of year 3. For each game, you record how many people are dead worldwide (in millions) at the end of year 4. You would like to assess the long run performance of this strategy, as measured by body count.



1. [3pt] Describe the shape, spread, and center of the distribution of the number of deaths in the 60 games in the sample.

   The distribution is right-skewed, with the bulk of the data between 882.999 and 1421.223. The distribution is centered around 1035.847, but, due to the skew, has a mean of 1227.786.

2. [2pt] Is this an experiment or an observational study? Explain why.
   This is an observational study. We are not assigning treatments to sampled units, rather, we are observing the pre-existing characteristics of those units.

3. [1pt] What type of variable is this?

   This is a quantitative variable

4. [1pt] Describe the population (in words)
   The population is the long-run average number of people dead from BubonicEbola at the end of year 4.

5. [1pt] Is it possible to collect data on every item in the population? If yes, how would you do it? If no, why not?

It is not possible to collect data on every item in the population because we cannot collect data on all future games which might be played with this strategy.

6. [1pt] Can you use simulation-based inference for this data? Why or why not?

No, we cannot use simulation-based inference for this data. We do not have knowledge about the entire population of potential outcomes.

7. [2pt] What are the validity conditions for theory-based inference on the type of data you selected in question 3?

We must have either a symmetric distribution OR at least 20 observations and a distribution that is not too strongly skewed.

8. [2pt] Are these validity conditions met? Why or why not?

Our distribution is not symmetric, so we must instead have at least 20 observations (we have 60) and a distribution that is not too strongly skewed. The distribution is skewed, but it is not skewed enough to be a problem for the use of theory-based inference.

9. [1pt] What formula would you use to calculate the standard deviation of the sample mean?

The standard deviation of the sample mean is $SE(\overline{x})$, and is calculated as $\frac{s}{\sqrt{n}}$.

10. [1pt] Calculate the standard deviation of the sample mean for your sample.

$s = 572.172$ and $n = 60$. Thus, $SE(\overline{x}) = 572.172/7.746 = 73.867$.

11. [2pt] What formula would you use to calculate an approximately 95% confidence interval for the population mean number of fatalities due to BubonicEbola using your strategy? Write it using the appropriate mathematical symbols for this type of variable.

$statistic \pm 2 \times SE(statistic) = \overline{x} \pm 2 \times \frac{s}{\sqrt{n}}$

12. [1pt] Calculate the approximately 95% confidence interval for the population mean number of BubonicEbola fatalities at the end of year 4. Show your work. If you do not have a value for a component of the formula, define a variable for that value and show as much work as you can.

$\overline{x} \pm 2\frac{s}{\sqrt{n}} = 1227.786 \pm 2\frac{572.172}{\sqrt{60}} = 1227.786 \pm 2(73.867) = (1080.052, 1375.52)$

13. [3pt] Interpret your interval from 12 in the context of the problem. If you did not get an interval from 13, use (a, b) as your values and interpret as normal.

I am 95% confident that the average number of deaths due to BubonicEbola at the end of year 4 is between 1080.052 and 1375.52.

Comparing Infection Strategies

_____/27

When evolving a disease in Plague, Inc., you have multiple different upgrades you can focus on - Symptoms, Transmission, and Abilities. Symptoms increase the likelihood of infection, while Transmission upgrades focus on how the disease is spread (e.g. through animal contact, blood transmission, air, or in humid environments). You decide to compare strategies, playing 60 games in total. You will randomly sample 30 numbers from 1 to 60; games with sampled numbers will use only transmission-focused upgrades to increase the spread of the disease, while games with numbers not in the random sample will use only symptom upgrades to increase infectivity. For each game, you will play 3 in-game years, and at the end of that period, you will assess whether the disease has spread so that at least 50% of the world is infected (a success). Is focusing on one type of upgrades more likely to infect at least 50% of the world in 3 years than focusing on the other type of upgrades?

Strategy	Killed >= 50%	Killed <50%	Total
Symptom	12	18	30
Transmission	17	13	30

1. [1pt] Is this an observational study or an experiment?

   This is an experiment, there is random assignment of strategies to games.

2. [1pt] What is the research question? (Please state in the form of a question, with appropriate punctuation)

   Is the success rate for the transmission upgrade strategy the same as the success rate for the symptom upgrade strategy?

3. [1pt] Are the variables categorical or quantitative?
   The variable (More than 50% of the world infected, or not) is categorical/binary. The group variable (Strategy) is also categorical.

4. [4pt] If you were using a hypothesis test to answer the research question,

   - What would your null hypothesis be in words?
     The success (more than 50% of the world infected) rate is the same for the two strategies.

   - What is the null hypothesis, in symbols? Be sure to use the appropriate symbols for the variable type you identified in 2!

     $H_0 : \pi_S = \pi_T$ *or* $\pi_T - \pi_S = 0$

   - What is the alternative hypothesis, in words?

     The success rate is not the same for the two strategies.

   - What is the alternative hypothesis, in symbols?

     $H_A : \pi_S \neq \pi_T$ *or* $\pi_T - \pi_S \neq 0$

5. [1pt] What is the relevant statistic for this problem? What is the population equivalent? Be sure to use the correct symbols.
   Population version: $\pi_T - \pi_S$
   Sample version: $\hat{p}_{\cdot T} - \hat{p}_S$

6. [2pt] What are the validity conditions for theory-based inference which are applicable to this problem? Are they met? Why or why not?

Each group must have both 10 successes and 10 failures. The validity conditions are met - there are 12 and 17 successes, and 18 and 13 failures; all quantities are greater than 10.

7. [2pt] Calculate the sample proportion for each group; using these values, calculate the value you identified in 5.

$\hat{p}_S = 12/30 = 0.4000$
$\hat{p}_T = 17/30 = 0.5667$
$\hat{p}_T - \hat{p}_S = 0.5667 - 0.4000 = 0.1667$

8. [2pt] What formula should you use to calculate the standard error of the quantity in 5, if you were to conduct a theory-based hypothesis test? If the formula uses symbols not defined in previous questions, please define those symbols as well. (Hint: Be sure you use the standard error formula for hypothesis tests, not confidence intervals! You may have to find this formula within a formula that is given to you.)

$$\hat{p} = \frac{n_T\hat{p}_T + n_S\hat{p}_S}{n_T + n_S} \qquad SE(\pi_T - \pi_S) = \sqrt{\hat{p}(1-\hat{p})\left(\tfrac{1}{n_T} + \tfrac{1}{n_S}\right)}$$

9. [2pt] What formula should you use to calculate the standard error of the quantity in 5, if you want to construct a confidence interval? (Hint: This is not the same formula as you used in question 7, and you may have to find it inside a different formula.)

$$SE(\hat{p}_T - \hat{p}_S) = \sqrt{\frac{\hat{p}_T(1-\hat{p}_T)}{n_T} + \frac{\hat{p}_S(1-\hat{p}_S)}{n_S}}$$

10. [3pt] Calculate a 2SD confidence interval for the difference in shiny encounter rates between the two species; be sure to use correct notation and show your work.

$$SE = \sqrt{\frac{17/30(13/30)}{30} + \frac{12/30(18/30)}{30}} = 0.12722$$

$$\hat{p}_T - \hat{p}_S \pm 2 \times SE = 0.16667 \pm 2(0.12722) = (-0.0878, 0.4211)$$

11. [4pt] Interpret your interval in the context of the problem. If you did not get an answer for 10, use the (wrong) interval (0.1503, 0.3201).

We are 95% confident that the difference in the proportion of games with more than 50% of the world infected using the transmission and symptom only strategies is between -0.0878 and 0.4211.

12. [2pt] Given your answer in 10, what can you conclude about the value of the standardized statistic? Explain your conclusion.
Note: Do not calculate the standardized statistic; instead, base your explanation on the confidence interval you interpreted.

As the interval contains 0, the magnitude of z will be less than 2.

13. [2pt] Given your answers in 10 and 11, would you reject or fail to reject the hypothesis H_0 : the shiny encounter rates between the two species are the same?

As the interval contains 0, it is plausible that the two strategies have equal probability of infecting at least 50% of the world in 3 years. I fail to reject H_0.