STAT22000 Summer 2020 Homework 13 Solutions

Problems to Turn In: due **5 pm on Thursday, July 23, on Gradescope**.

1. Refer to Exercise 6.8 on p.313 in the textbook. Instead of doing the two parts therein, determine if the following statements are true or false, and explain your reasoning.

   (a) We are 95% confident that 63% to 69% of American adults in this sample think licensed drivers should be required to retake their road test once they turn 65.

   (b) We are 95% confident that 63% to 69% of American adults think licensed drivers should be required to retake their road test once they turn 65.

   (c) If we take many random samples of 1018 American adults, and for each sample, calculated the percentage who think licensed drivers should be required to retake their road test once they turn 65. 95% of those sample percentages will be between 63% and 69%.

   (d) The margin of error at a 99% confidence level would be higher than 3%.

   Answer: *[4pts in total, 1pt each. It's wrong if the reason is wrong.]*

   (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion.

   (b) True. This is the correct interpretation of the confidence interval, which can be calculated as $0.66 \pm 0.03 = (0.63, 0.69)$.

   (c) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion of another sample. It does not tell us what we might expect to see in another random sample.

   (d) True. As the confidence level increases, the margin of error increases as well.

   ---

2. This problem is about the data set in Lab 8:

   `http://www.stat.uchicago.edu/~yibi/s220/labs/lab08.html`

   which is a random sample of 1000 birth records in the state of North Carolina. This time we are interested in the percentage of babies that had low birth weight (below 2500 gram or 5 pounds 8 ounces).

   (a) Find the number of babies in the 1000 birth records that had low birth weights using the R commands below.

   ```
   nc = read.csv("https://www.openintro.org/stat/data/csv/ncbirths.csv")
   library(mosaic)
   tally(~lowbirthweight, data=nc)
   ```

   Estimate the percentage of babies in North Carolina that had low birth weight and calculate a 95% confidence interval for it.

   Answer: *[3pts = 1pt for the estimate + 1pt for 1.96 + 1pts for the SE]* Using the R codes provided, we can see that 111 of the 1000 sampled birth records had a low birth weight. The estimated percentage is $\widehat{p} = 111/1000 = 11.1\%$. The 95% CI for the population percentage is

   $$\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0.111 \pm 1.96\sqrt{\frac{0.111(1-0.111)}{1000}} \approx 0.111 \pm 0.0195 \approx (0.0915, 0.1305) = (9.15\%, 13.05\%).$$

(b) If we want to reduce the margin of error for the 95% confidence level to 1.5%, how large the size of a sample would you recommend?

Answer: *[3pts]* The margin of error for a 95% CI is $z^*\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ where $z^* = 1.96$. To reduce the margin of error to 1.5% = 0.015, the sample size n should be at least

$$\text{margin of error} = 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq 0.015 \quad \Rightarrow \quad n \geq \left(\frac{1.96}{0.015}\right)^2 \widehat{p}(1-\widehat{p})$$

As the sample proportion $\widehat{p}$ is unknown before the new data are collected, we need to guess the value of $\widehat{p}$. A reasonable guess is the $\widehat{p} = 0.111$ from the previous sample. As we are not sure what the true p is, one can also try other p in the CI (0.0915, 0.1305) in the previous part, or use the conservative guess (but less reasonable) $\widehat{p} = 0.5$ o since $\widehat{p}(1-\widehat{p})$ achieves the greatest value when $\widehat{p} = 0.5$.

$$n \geq \left(\frac{1.96}{0.015}\right)^2 \widehat{p}(1-\widehat{p}) = \begin{cases} 1685 & \text{if } \widehat{p} = 0.111 \\ 1938 & \text{if } \widehat{p} = 0.1305 \\ 4269 & \text{if } \widehat{p} = 0.5 \end{cases}$$

[All answers above are acceptable.]

(c) According to CDC (`https://www.cdc.gov/nchs/fastats/birthweight.htm`), 8.17% of the babies born in 2016 in the U.S. had a low birth weight. Test whether the state of North Carolina had a *higher* percentage of babies having low birth weight than the nationwide percentage 8.17%. Please specify the null and alternative hypotheses, report the test statistic and the P-value, and make a conclusion using 0.01 significance level.

Answer: *[5pts = 1pt for the hypotheses + 2pts for the z-statistic + 1pt for the one-sided P-value + 1pt for the conclusion.*
Take 1pt off if SE is calculated as $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \sqrt{\frac{0.111\times0.889}{1000}}$ *rather than* $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.0817(1-0.817)}{1000}}$.
Take 0.5pt off if a two-sided P-value is calculated rather a one-sided P-value.]
The hypotheses are H$_0$: $p = 0.0817$ and H$_a$: $p > 0.0817$, where p is proportion of babies born in North Carolina had a low birth weight *[if stated in this way, must explain what p is or 0.5pt off]*. Alternatively, the hypotheses can also be stated verbally as

H$_0$: The percentage of babies born in North Carolina that had a low birth weight is 8.17%, same as the nationwide percentage.

H$_a$: The percentage of babies born in North Carolina that had a low birth weight is over 8.17%, higher than the nationwide percentage.

[0.5pt off if H_a is two-sided. 1pt off if the hypotheses are stated in terms of the sample percentage $\widehat{p}$.]
The test statistic is $z = \dfrac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{0.111 - 0.0817}{\sqrt{0.0817(1-0.0817)/1000}} \approx \dfrac{0.0293}{0.00866} \approx 3.38$.
The upper one sided P-value is $P(Z > 3.38) \approx 0.00036$.

```
> pnorm(3.38,lower.tail=F)
[1] 0.0003624291
```

Since the p-value is lower than the significance level 0.01, we reject H$_0$. The data provide strong evidence that the percentage of babies in NC that had a low birth weight is higher than the nationwide percentage 8.17%.

3. This problem is a continuation of Problem 2. This time we are interested in testing whether babies born to smoking mothers had a percentage of having low birth weight (below 2500 gram or 5 pounds 8 ounces) than those born to nonsmoking mothers

   (a) Make a two-way table cross classify the 1000 babies by whether they had low birth weights and whether the mother smoked using the R commands below.

```
nc = read.csv("https://www.openintro.org/stat/data/csv/ncbirths.csv")
library(mosaic)
tally(~habit+lowbirthweight, data=nc)
```

Answer: *[0pt]* Using the R command given, we get the following output.

```
           lowbirthweight
habit         low not low
  nonsmoker    92     781
  smoker       18     108
  <NA>          1       0
```

We see there is one birth record that the information on mother's smoking habit is missing (`NA`). The two-way table is

	Birth Weight?		
	Low	Not Low	Total
Non-smoker	92	781	873
Smoker	18	108	126
NA	1	0	0
Total	111	889	1000

Or if we ignore the birth record with missing value, the two-way table is

	Birth Weight?		
	Low	Not Low	Total
Non-smoker	92	781	873
Smoker	18	108	126
Total	110	889	999

   (b) Test if babies born to smoking mothers had a higher percentage of having low birth weights than those born to non-smoking mothers. State the hypotheses. Report the test statistic and the P-value. What is your conclusion at significance level $\alpha = 0.05$? (Please ignore the observation that the mother's smoking habit is missing.)

Answer: *[4pts = 1pts for the hypotheses + 2pt for the z-statistic + 0.5pt for the one-sided P-value + 0.5pt for the conclusion. It's okay if the success-failure condition is not checked though it should be checked.]*

	Birth Weight?				
	Low	Not Low	Total		
Non-smoker	92	781	873	$\Rightarrow n_n = 873,$	$\widehat{p}_n = 92/873 \approx 0.1054$
Smoker	18	108	126	$\Rightarrow n_s = 126,$	$\widehat{p}_s = 18/126 \approx 0.1429$

The hypotheses are: H$_0$: $p_n = p_s$ and H$_a$: $p_n < p_s$, where p_n and p_s are respectively the proportion of babies born to non-smoking mothers and smoking mothers that had a low birth weight. [*0.5pt off if not explaining p_t and p_c.*]

The hypotheses can also be stated verbally as

H$_0$: The percentages of babies that had a low birth weight among those born to smoking mothers and among those born to nonsmoking mothers are equal. was identical to the percentage

H$_a$: The percentages of babies that had a low birth weight is higher among those born to smoking mothers than among those born to nonsmoking mothers.

If H$_0$ is true, the pooled estimate of the common p is

$$\widehat{p} = \frac{92+18}{873+126} = \frac{110}{999} \approx 0.1101.$$

We can safely use a large sample z test since the number of successes and failures in the two samples:

$$\begin{aligned} n_n\hat{p} &= 873 \times \frac{110}{999} \approx 96.1, \quad & n_n(1-\hat{p}) &= 435(1-\frac{110}{999}) \approx 776.9, \\ n_s\hat{p} &= 126 \times \frac{110}{999} \approx 13.9, \quad & n_s(1-\hat{p}) &= 435(1-\frac{110}{999}) \approx 112.1 \end{aligned}$$

are all greater than 10.
The SE is

$$\text{SE} = \sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_n} + \frac{1}{n_s}\right)} = \sqrt{0.1101(1-0.1101)\left(\frac{1}{873} + \frac{1}{126}\right)} \approx 0.0298.$$

and the z-statistic is

$$z = \frac{\widehat{p}_s - \widehat{p}_n}{\text{SE}} = \frac{(18/126) - (92/873)}{0.0298} \approx 1.256.$$

[*Take 1pt off if SE is calculated as* $\sqrt{\frac{0.1054(1-0.1054)}{873} + \frac{0.1429(1-0.1429)}{126}} \approx 0.0329$*, and the z-statistic becomes 1.14.*]
The one-sided P-value is $P(Z > 1.256) \approx 0.10456$.

```
> pnorm(1.256, lower.tail=F)
[1] 0.104558
```

[*Take 0.5pt off if a two-sided P-value is calculated rather than a one-sided P-value.*] The percentages of babies that had a low birth weight is not significantly higher among those born to smoking mothers than among those born to nonsmoking mothers.

---

4. (Revision of Exercise 6.26) A 2010 Pew Research foundation poll indicates that among 1099 college graduates, 33% watch The Daily Show. Meanwhile, 22% of the 1110 people with a high school degree but no college degree in the poll watch The Daily Show. A 99% confidence interval for $p_{\text{college grad}} - p_{\text{HS}}$, where p is the proportion of those who watch The Daily Show, is $(0.06, 0.16)$.

   (a) Verify that the 99% CI for $p_{\text{college grad}} - p_{\text{HS}}$ is about $(0.06, 0.16)$. Show your calculation.

   (b) Determine if the following statements are true or false, and explain your reasoning if you identify the statement as false.

      i. There was strong evidence that $p_{\text{college grad}} > p_{\text{HS}}$.

ii. 99% of random samples of 1,099 college graduates and 1,110 people with a high school degree but no college degree will yield differences in sample proportions between 6% and 16%.

iii. A 95% confidence interval for $p_{\text{college grad}} - p_{\text{HS}}$ would be wider.

iv. A 99% confidence interval for $p_{\text{HS}} - p_{\text{college grad}}$ is $(-0.16, -0.06)$.

Answer:

(a) *[2pts. It's okay if the success-failure condition is not checked though it should be checked.]*

We can safely use a large sample CI since the number of successes and failures in the two samples:

$$1099 \times 0.33, \quad 1099 \times (1-0.33), \quad 1110 \times 0.22, \quad 1110 \times (1-0.22)$$

are all greater than 10.

An approximate 99% CI for $p_{\text{college grad}} - p_{\text{HS}}$ is $\boxed{\text{estimate} \pm 2.58\,\text{SE}}$ where

$$\text{estimate} = \widehat{p}_{\text{college grad}} - \widehat{p}_{\text{HS}} = 0.33 - 0.22 = 0.11$$
$$\text{SE} = \sqrt{\frac{0.33(1-0.33)}{1099} + \frac{0.22(1-0.22)}{1110}} = 0.01886$$

So the 99% CI is $0.11 \pm 2.58 \times 0.01886 \approx 0.11 \pm 0.049 \approx 0.11 \pm 0.05 = (0.06, 0.16)$. *[Take 1pt off if SE is calculated as* $\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{1099} + \frac{1}{1110}\right)} \approx 0.019$ *where* $\widehat{p} = \frac{0.33\times1099+0.22\times1110}{1099+1110} \approx 0.275$.*]*

(b) *[4pts in total. 1pt each. It's wrong if the reason is wrong.]*

i. True.

ii. False. The confidence interval is for enclosing a population parameter, not a sample statistic.

iii. False. As the confidence level decreases the width of the confidence interval decreases.

iv. True.
