Discussion 8

Hanying Jiang

Please download the Discussion 9 materials, including the data and rmd file.

Hw6 2a(iii) and 2b(iii)

We have collected a random sample of size n with sample mean $\bar{X}$. We also know that the population SD is σ. Assume the population is normal or n is large enough for applying CLT. We have constructed a 90% CI for the population mean μ: $[\mathrm{CI}_{\mathrm{lower}}, \mathrm{CI}_{\mathrm{upper}}]$. Besides, we want to perform a hypothesis test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ with $\alpha = 0.1$. Then

If $\mu_0 \in [\mathbf{CI}_{\mathbf{lower}}, \mathbf{CI}_{\mathbf{upper}}]$, then we must have p-value $> \alpha$ and we won't reject the null, and vice versa.

Exercise 1.

For question (a)-(c), run the given code and check the output. Raise your hand to let me know if you have difficulty getting the results.

Question(d): perform the one-sample Z test for population proportion. Check details in slides for 2023/10/19.

Question(e): construct CI for difference between proportions. Check details in slides for 2023/10/31.

Question

Possible relationships between mothers' characteristics and birth outcomes have been studied extensively. Using the `lbw` data set, consider what inference tools we could apply to address the following questions. This data was collected as part of a larger study at Bayside Medical Center, Springfield, Massachusetts in 1986. It contains information on 189 births to women that were seen in the obstetrics clinic.

a. Load the lbw data into your environment by reading in the CSV file (lbw.csv) to the variable lbw. Make sure the csv file is in the same folder as your .Rmd file and that folder is set as your working directory.

   I have included vectors of the relevant data at the bottom of this page if you are having issues importing and would rather just define vectors of data.

```
lbw <- read.csv("lbw.csv", header=TRUE)
```

b. Run `str(lbw)`. Confirm that you see the 11 variables types. We will be focusing on the variables `smoke` (Did the mother smoke?) and `low` (Did the baby have low birth weight?). Set both of those variables to be categorical using the as.factor() command. Check that they have set correctly with str().

```
str(lbw)
lbw$smoke <- as.factor(lbw$smoke)
lbw$low <- as.factor(lbw$low)
str(lbw)
```

c. Since we will be focusing on the variables `smoke` and `low`, use the code below to keep only the relevant columns of data from the "lbw.csv" data set to a smaller data set named smoke.lbw. The table function will build a 2X2 contingency table for this data.

```
smoke.lbw <- lbw[,c(2,3)] # Keep only the second and third columns
table(smoke.lbw)
```

```
##     smoke
## low  0  1
##   0 86 44
##   1 29 30
```

d. A recent study reported that 7.1% of mothers smoked during pregnancy (https://reproductive-health-journal.biomedcentral.com/articles/10.1186/s12978-019-0807-5). Conduct a hypothesis test at an $\alpha = 0.05$ level to determine if there is evidence that the population from which this sample was selected has a different proportion of pregnant mothers who smoke.

(i) Define parameters and write the null and alternative hypotheses.

(ii) Run the following code to compute the number of patients in the sample of 189 who smoked and who did not. Or read this value off of the table you found above.

```
sum(smoke.lbw$smoke==1) #74 smokers
```

```
## [1] 74
```

```
sum(smoke.lbw$smoke==0) #115 nonsmokers
```

```
## [1] 115
```

(iii) Check the sample size assumptions necessary to use a z test.

(iv) Calculate the observed test statistic.

(v) Compute a p-value and make the appropraite conclusion.

e. Construct a 90% CI for $\pi_s - \pi_n$, the difference in proportion of babies classified as low birth weight (`lbw = 1`) between the babies with mothers who smoked during pregnancy (`smoke = 1`) and mothers who did not smoke (`smoke = 0`). Interpret your findings.

Point Estimate

Standard Error

Critical Value

Margin of Error

Confidence Interval

f. When considering the relationship between mother's smoking and low birth weight of their babies, what other variables might be confounding?

Solution

d.

A recent study reported that 7.1% of mothers smoked during pregnancy. Conduct a hypothesis test at an $\alpha = 0.05$ level to determine if there is evidence that the population from which this sample was selected has a different proportion of pregnant mothers who smoke.

(i) Define parameters and write the null and alternative hypotheses.

Let π_s be the proportion of smoking mothers in the population from which the sample was taken.

$H_0 : \pi_s = 0.071$ vs. $H_A : \pi_s \neq 0.071$.

(ii) Run the following code to compute the number of patients in the sample of 189 who smoked and who did not. Or read this value off of the table you found above.

```
sum(smoke.lbw$smoke==1) #74 smokers
```

```
## [1] 74
```

```
sum(smoke.lbw$smoke==0) #115 nonsmokers
```

```
## [1] 115
```

(iii) Check the sample size assumptions necessary to use a z test.

Requirement: $n\pi_0 \geq 5$ and $n(1-\pi_0) \geq 5$.

$189 * 0.071 = 13.419 \geq 5$.

$189 * (1 - 0.071) \geq 5$.

So the sample size is large enough to apply CLT, and $\hat{p}$ is approximately normal.

(iv) Calculate the observed test statistic.

Sample proportion: $\hat{p} = \frac{74}{189} = 0.3915344$. The observed test statistic:

$$Z_{\text{obs}} = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.3915344 - 0.071}{\sqrt{0.071 * (1 - 0.071)/189}} = 17.2116.$$

(v) Compute a p-value and make the appropriate conclusion.

```
2 * pnorm(-17.2116) # two sided
```

```
## [1] 2.173429e-66
```

Or

```
2 * (1 - pnorm(17.2116))
```

```
## [1] 0
```

e.

Construct a 90% CI for $\pi_s - \pi_n$, the difference in proportion of babies classified as low birth weight (`lbw = 1`) between the babies with mothers who smoked during pregnancy (`smoke = 1`) and mothers who did not smoke (`smoke = 0`). Interpret your findings.

Point Estimate

We use the sample proportion as the point estimate of the population proportion:

$$\hat{p}_s - \hat{p}_n = \frac{30}{74} - \frac{29}{115} = 0.1532315.$$

Standard Error

$$\begin{aligned} SE(\hat{p}_s - \hat{p}_n) &\approx \sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{n_s} + \frac{\hat{p}_n(1-\hat{p}_n)}{n_n}} \\ &= \sqrt{\frac{\frac{30}{74} * \frac{44}{74}}{74} + \frac{\frac{29}{115} * \frac{86}{115}}{115}} \\ &= 0.06998073. \end{aligned}$$

Critical Value

Since $n_s\hat{p}_s, n_s(1-\hat{p}_s), n_n\hat{p}_n, n_n(1-\hat{p}_n)$ $(30, 44, 29, 86)$ are all at least 5, we can apply CLT and approximate $\hat{p}_s - \hat{p}_n$ with normal distribution. Thus, we can use Z CI and the critical value is

$$Z_{0.95} = 1.644854.$$

```
qnorm(0.95)
```

```
## [1] 1.644854
```

Margin of Error

$$Z_{0.95} * SE(\hat{p}_s - \hat{p}_n) = 1.644854 * 0.06998073 = 0.1151081.$$

Confidence Interval

$$\text{point estimate} \pm \text{margin of error} = 0.15323 \pm 0.1151081 == (0.0381, 0.2683).$$

Interpretation

This CI suggests that the proportion of babies born classified as low birth weight is different than the population of babies with mothers who smoked during pregnancy, since the confidence interval does not contain 0.

f

When considering the relationship between mother's smoking and low birth weight of their babies, what other variables might be confounding?

There are lots of other variables that might have a relationship with both the mother's smoking and the birth weight of the baby: number of prenatal appointments attended, other health metrics of mother. We would need more advanced analysis strategies to account for those other possible confounding variables.