Problem Set 6 - Hypothesis Testing with 1 and 2 Means

2024-01-08

Part 1: Concepts and Calculations

For the first set of problems, you'll be calculating the results by hand (sort of - you can use R as your giant calculator). When appropriate, show your work in a chunk of R code.

Question 1

The survey of 300 college students found that the average semester expenditure was $350 with a standard deviation of $78. At the same time, campus administration has done an audit of required course materials and claims that the average cost of books and supplies for a single semester should be no more than $340. In other words, the administration is saying the population mean value is $340.

a. State a null and alternative hypothesis to test the administration's claim. Did you use a one or two tailed alternative hypothesis? Explain your choice.

Ho: m = 340 Ha: m > 340

For this test, I used a one-tailed alternative hypothesis because the two options that we want to verify are that the average semester expenditure for books is either equal to 340, or bigger than 340 (which is our alternative). As a result, we should perform a one-tailed test.

b. Test the null hypothesis and discuss the findings. Show all calculations.

```r
#Get z score for alpha = 0.05 (from z score table)
-1.65
```

```
## [1] -1.65
```

```r
#Calculate z
(350-340)/78
```

```
## [1] 0.1282051
```

```r
#Compare |z| with c.v: since |0.12|< |1.65|
```

As a result, we fail to reject the null hypotheses, which means that we agree that the average value of semester expenditures is probably equal to 340.

Question 2

In response to the student survey that we keep talking about, a potential donor wants to provide campus bookstore gift certificates as a way of defraying the cost of books and supplies. In consultation with the student government leaders, the donor decides to prioritize first and second year students for this program because they think that upper-class students do not spend as much on books and supplies. Before finalizing the decision, the student government wants to test whether there really is a difference in the spending patterns of the two groups of students. What are the null and alternative hypotheses for this problem? Explain.

For this problem, I would set Ho: lower classmen expenditures = upper classmen expenditures Ha: lower classmen expenditures =/ upper classmen expenditures By doing this two-tailed test, we could conclude whether or not lower classmen seem to be spending more than upper classmen, and how to distribute the money so that both groups could benefit equally.

Part 2: R Problems

For these problems, you should use the `county20large` data set to examine how county-level educational attainment is related to COVID-19 cases per 100k population. You need to load the following libraries: `dplyr, Hmisc, gplots, descr, effectsize.` Remember to write a line in your markdown document that tells R where to get the data from (hint: it’s in the data folder).

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(gplots)
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(descr)
library(effectsize)
load("data/county20large.rda")
```

Question 1

The first thing you should do is take a sample of 500 counties from the `counties20large` data set and store that sample in a new data set, `covid500` using the command listed below.

```
set.seed(1234)
#create a sample of 500 rows of data from the county20large data set
covid500 <- sample_n(county20large, 500)
```

The `sample_n` command samples rows of data from the data set so we now have 500 randomly selected counties with data on all of the variables in the data set. The dependent variable in this assignment is `covid500$cases100k_sept821` (cumulative COVID-19 cases per 100,000 people up to Sept 8 2021) and the independent variable is `covid500$postgrad`, the percent of adults in the county with a post-graduate degree, and the expectation that case rates are lower in counties with relatively high levels of education than in other counties.

Question 2

Transform `covid500$postgrad` into a two-category variable with a roughly equal amount of counties in each category. Store this variable in a new object named `covid500$postgrad2` and label the categories "Low Education" and "High Education." The generic format is `data$newvariable <- cut2(data$oldvariable, g=number of groups)`. If you are unclear about how to do this, go back to Chapter 4 for a refresher. Produce a frequency table for `covid500$postgrad2` to check on the transformation.

```
covid500 <- sample_n(county20large, 500)
covid500$postgrad2 <- cut2(covid500$postgrad, g=2)
levels(covid500$postgrad2)<-c("Low Education","High Education")
table(covid500$postgrad2)
```

```
##
##  Low Education High Education
##            257            242
```

```
freq(covid500$postgrad2, plot=F)
```

```
## covid500$postgrad2
##                  Frequency Percent Valid Percent
## Low Education          257    51.4          51.5
## High Education         242    48.4          48.5
## NA's                     1     0.2
## Total                  500   100.0         100.0
```

Question 3

State a null and an alternative hypothesis for this pair of variables.

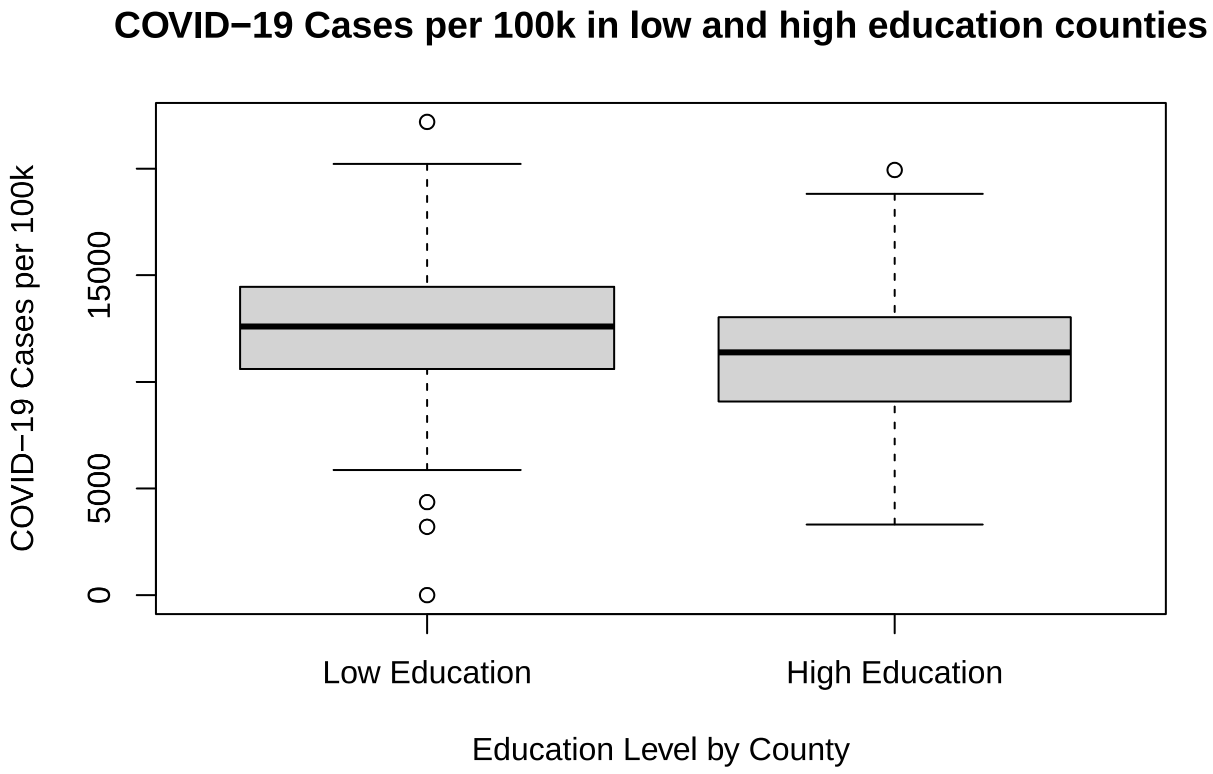
Ho: loweduc = higheduc Ha: loweduc =/ higheduc

Question 4

Use the `compmeans` command to estimate the level of COVID-19 cases per 100k in low and high education counties. Describe the results. What do the data and boxplot tell you? Make sure to use clear, intuitive labels for the boxplot and make specific references to the group means.

```
compmeans(covid500$cases100k_sept821, covid500$postgrad2,plot=T,
 xlab="Education Level by County",
 ylab="COVID-19 Cases per 100k",
 main="COVID-19 Cases per 100k in low and high education counties")
```

```
## Warning in compmeans(covid500$cases100k_sept821, covid500$postgrad2, plot = T,
## : 1 rows with missing values dropped
```



```
## Mean value of "covid500$cases100k_sept821" according to "covid500$postgrad2"
##                     Mean   N Std. Dev.
## Low Education  12561.40 257  3134.551
## High Education 11159.39 242  2975.240
## Total          11881.47 499  3134.739
```

From our data we can gather that counties with lower values of education, on average had more COVID-19 cases per 100k of population, around 12746.07 cases. On the other hand, counties with high values of education had lower averages of cases per 100k citizens, around 11433.21 cases, for a difference of around 1313 cases, on average, between low and high education counties. From the boxplot, we can observe that low education counties had a few high outliers, while high education counties had one very low value outlier. The fact that both counties' medians are situated more or less halfway between their 25th and 75th percentiles, shows that the results are not significantly skewed.

Question 5

Conduct a t-test for the difference in COVID-19 rates between low and high education counties. Interpret the results.

```
t.test(covid500$cases100k_sept821~covid500$postgrad2)
```

```
##
##  Welch Two Sample t-test
##
```

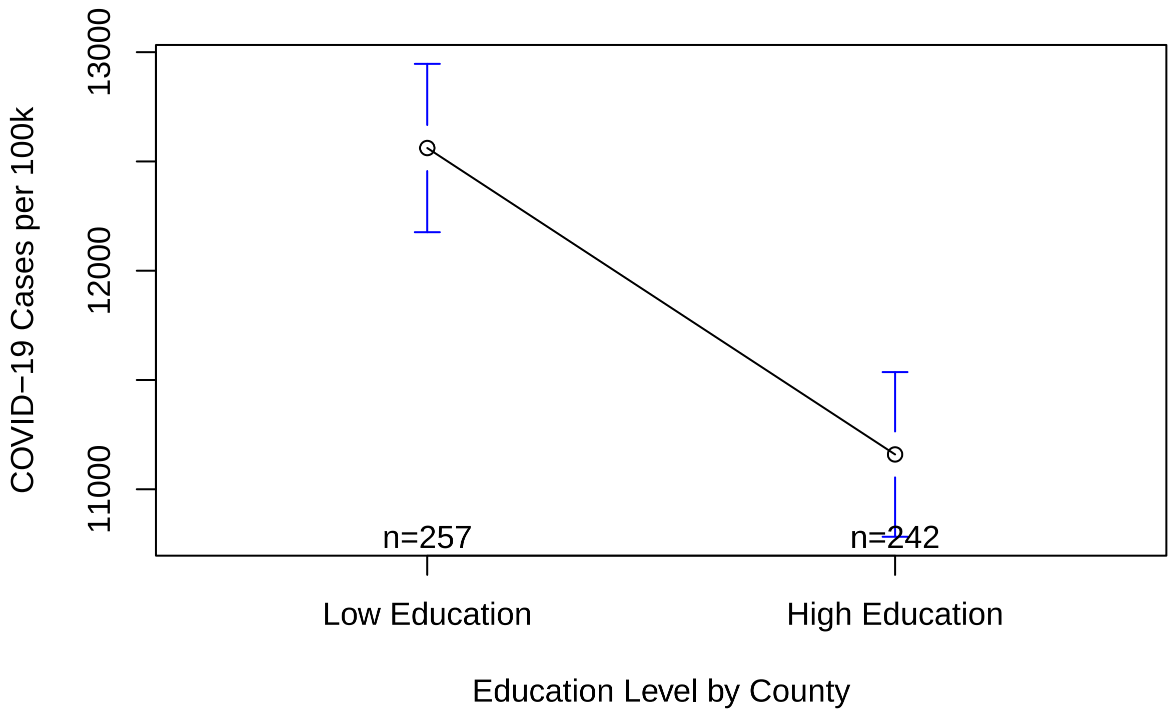
```
## data:  covid500$cases100k_sept821 by covid500$postgrad2
## t = 5.1259, df = 496.97, p-value = 4.248e-07
## alternative hypothesis: true difference in means between group Low Education and group High Educatio
## 95 percent confidence interval:
##   864.626 1939.399
## sample estimates:
##  mean in group Low Education mean in group High Education
##                     12561.40                     11159.39
```

Since the p-value > 0.05, we fail to reject the null hypothesis and we can conclude that the difference in cumulative COVID-19 cases per 100,000 people up to Sept 8 2021 between low education and high education counties is equal to zero, meaning that the averages of the two are very similar.

Question 6

Add a means plot (`plotmeans` command) and Cohen's D (`cohens_d` command) and discuss what additional insights they provide.

```
plotmeans(covid500$cases100k_sept821~covid500$postgrad2,
n.label=T,
ylab="COVID-19 Cases per 100k",
xlab="Education Level by County")
```



One additional insight provided by plotmeans are the error bars, which represent the confidence intervals around each of the two subgroup means. The vertical distance between the two group means, and the fact that neither of teh confidence intervals overlaps with each other, shows us that the two counties are statistically different from one another.

```
cohens_d(covid500$cases100k_sept821~covid500$postgrad2)
```

```
## Cohen’s d |       95% CI
## ------------------------
## 0.46      | [0.28, 0.64]
##
## - Estimated using pooled SD.
```

Cohen’s D measures the size of the effect, or the difference between the two group means relative to the size of the pooled standard deviation. In this case, Cohen’s D is 0.38, which, according to the Table reported on page 246 of the book, can be interpreted as the level of education by counties to have a quite small effect on COVID-19 cases per 100k of population.