Public Policy 529
Fall 2023 Problem Set #7

Francisco Brady

2023-11-06

```r
library(haven)
library(dplyr)
library(questionr)
anes <- read_dta('anes2020subset.dta')
```

Due Monday, November 6th, end of day

1. Facing claims that city police were engaging in racial profiling, the city of Grand Rapids hired a consulting firm to perform a study on traffic stops in the city. The results of this study were released in April 2017, and the consulting firm's report is posted on Canvas in the Problem Sets folder. In short, the study found that Black motorists were stopped at "close to twice the rate that would be expected given their presence in the traffic." It is useful to examine the study's methodology. First, the consulting firm collected benchmark data on the race of drivers at particular intersections in the city. Thus, for each location, we have a sample with information on the percentage (i.e. proportion) of drivers that are Black. Second, the consulting firm collected data from the police department on the race of people who were stopped near those same locations, providing a second sample that measures the proportion of drivers that are Black. Under the null hypothesis of no racial profiling, the percentages in these independent samples are the same.

(a) Earlier in the course, we talked about measurement. Examine how the consulting firm measured the benchmark data on the race of drivers (pp. 30-39 of the study). Assess the reliability and validity of this measurement strategy. Reliability has to do with consistency, and whether or not the measure can be replicated by others at a different time. The consulting firm picked 20 spots to conduct their analysis, not in a random way, but based on police activity, and a number of other factors.
The study does mention that they did their own tests of reliability across surveyors – the people who assigned race/ethnicity to the drivers during their benchmarking, and found that the raters assigned the same race to a specific driver 88% of the time, and were 92% accurate after gaining some experience, giving a rough estimate of how reliable their surveyors were. The consultants made efforts to cover all possible time periods except when there were few drivers at the selected locations. The surveys were conducted from August through November. It's not clear whether a similar test conducted during a different time of the year would yield different results. Overall it seemed like the consultants were aware of the ways in which their surveys could be affected by various factors, and took steps to mitigate the possibility that various outside factors would affect their results.

Validity has to with how well the scores from a measure represents the variable that it is intended to measure. The consultants set out to measure the difference between a benchmark measure – how frequently drivers at specific locations were from different racial, ethnic and gender groups, and a stop measure – how frequently drivers were stopped from each racial, ethnic, or gender group. The measure is intended to score the difference between the benchmark and the number of stops. In this respect it is a valid measure.

(b) According to the data (p. 56), at the corner of Alpine & Leonard, 13.8% of the 3,042 drivers were Black. Out of 487 traffic stops made in that vicinity, 27.5% of the drivers were Black. Construct a 95% confidence interval for the difference of proportions. Be sure to use the correct standard error for a confidence interval.
$\hat{\pi}_1 = .138$, $n_1 = 3042$, $\hat{\pi}_2 = .275$, $n_2 = 487$

$$ci = (\hat{\pi}_2 - \hat{\pi}_1) \pm z \cdot (se_{diff})$$

Assuming independent samples, and no assumption of equal variances:

$$se_{diff} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

$$se_{diff} = \sqrt{\frac{.138(.862)}{3042} + \frac{.275(.725)}{487}}$$

```
se_diff <- sqrt(
  ((.138*.862) / 3042) + ((.275*.725)/487)
  )
```

$$se_{diff} = 0.02117779$$

Using this in the ci formula:

$$ci = 0.137 \pm 1.96 \cdot (0.02117779)$$

So the 95% confidence interval would be between: (0.09549153, 0.1785085)

```
ci <- 0.137 - (1.96 *0.02117779)
ci <- 0.137 + (1.96 *0.02117779)
```

(c) Now perform a significance test ($\alpha = .01$) in which the null hypothesis is that there is no difference between the proportion of drivers who are Black and the proportion of traffic stops that involve Black drivers. Perform all the steps and report all relevant statistics.

i) assumptions: independent samples, large sample sizes. there is at least 10 cases in each category. proportions so using z tests.
ii) hypothesis: $H_0 : \hat{\pi}_2 - \hat{\pi}_1 = 0$, $H_a : \hat{\pi}_2 - \hat{\pi}_1 \neq 0$
iii) Critical value and test statistic: Critical value at $\alpha_{.01} \approx 2.32$
To calculate the test statistic, we first need the standard error using the baseline assumption of no difference. To find $\hat{\pi}$, we need to use the formula:

$$\hat{\pi} = \frac{\hat{\pi}_1 n_1 + \hat{\pi}_2 n_2}{n_1 + n_2} = \frac{(.138 \cdot 3042) + (.275 \cdot 487)}{3042 + 487} = 0.1571405$$

```
pi_hat <- (
  (.138 * 3048) + (.275*487)
  ) / (3042 + 487 )
pi_hat
```

```
## [1] 0.1571405
```

Utilizing $\hat{\pi}$ in our standard error calculation:

$$se_0 = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}$$

$$se_0 = \sqrt{\frac{0.1571405(1-.138)}{3042} + \frac{0.1571405(1-.138)}{487}}$$

```r
se_0 <- sqrt(((0.1571405 * .862) / 3042) + ((0.1571405 * .862) / 487) )
se_0
```

```
## [1] 0.01796302
```

To calculate the observed z statistic:

$$z = \frac{(\hat{\pi_2} - \hat{\pi_1}) - 0}{se_0}$$

$$z = \frac{(.275 - .138) - 0}{0.01796302}$$

```r
z <- ((.275 - .138) - 0)  / 0.01796302
z
```

```
## [1] 7.62678
```

iv) p-value: Our observed test statistic is 7.62678, so we know that our p-value is less than .000000287*2 = 0.000000574

v) conclude: Since the value of our test statistic exceeds the value of our z-statistic, we can reject the null hypothesis that the proportions are the same at a confidence level of 99%.

2. One important measure of development in a country is the rate of life expectancy. Suppose that, across a sample of democracies, the mean of this variable is 71.2 (n=77; s=10.2), and in a sample of non-democracies the mean is 64.8 (n=15; s=11.3).

(a) Are these independent or dependent samples? Explain. Yes, the samples are independent, because the life expectancy variable is measured in different samples. Further, the samples are taken from different countries. If the samples were from the same country, before and after a it became a democracy, they would be paired samples, and dependent.

(b) Do you think we should make the assumption that life expectancy rates have the same variance in the populations of democracies and non-democracies? Explain. We should not make the assumption that life expectancy has the same variance in democracies and non-democracies. There are many other factors that could contribute to the variance in life expectancy for democracies versus non-democracies, such as political stability, level of development, and differences in country wealth distributions. Another reason is because the standard deviation of democracies is 10.2, and in non-democracies it is 11.3. We would have to run an additional test to accept or reject that the difference between the standard deviations was in fact zero, but at first glance they imply a different variance for each sample. ($s_1^2 = 104.04, s_2^2 = 127.69$)

(c) Suppose we cannot assume that these samples come from populations with equal variances, find the standard error of the difference for mean years of life expectancy and estimate degrees of freedom using the shortcut method.

$$se_{diff} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

```
se_diff <- sqrt(
  ((11.3^2)/15) +
    ((10.2^2)/77)
  )
se_diff
```

```
## [1] 3.140674
```

$$se_{diff} = \sqrt{\frac{10.2^2}{77} + \frac{11.3^2}{15}} = \sqrt{1.351169 + 8.512667} = 3.140674$$

To estimate the degrees of freedom:

$$min(n_1 - 1, n_2 - 1) = min(77 - 1, 15 - 1) = 14$$

(d) Using the standard error you just calculated, perform a significance test for the difference of means (α = .05).

i) assumptions: not equal variance. Sample sizes are small, so assuming a normal distribution. Means so using a t statistic.
ii) hypothesis: $H_0 : \mu_2 - \mu_1 = 0$, $H_a : \mu_2 - \mu_1 \neq 0$.
iii) Critical value and test statistic: Critical value at $\alpha = .05$, with degrees of freedom $min(n_1-1, n_2-1) = 14$: $t_{.025} = 2.145$. To calculate the test statistic, we calculate the difference of means and use se_{diff} calculated above:
$$t = \frac{\bar{x_2} - \bar{x_1} - 0}{se_{diff}} = \frac{71.2 - 64.8 - 0}{3.140674} = 2.037779$$
iv) p-value: Our observed test statistic is 2.037779. Finding the closest t value at 14 degrees of freedom, and multiplying the confidence level ranges by 2: $t_{.050} < p < t_{.10}$.
v) Since absolute value of our test statistic **does not** exceed our critical value, so we cannot reject the null hypothesis that the difference between the two means is zero. Similarly, since the calculated p-value is greater than .05, we cannot reject the null hypothesis.

(e) Suppose instead that we can assume these samples come from populations with equal variances. What will be the standard error and degrees of freedom for the difference of means test? The standard error for equal variance pools the sample variances:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(77 - 1)10.2^2 + (15 - 1)11.3^2}{77 + 15 - 2}}$$

```
s <- sqrt(
  (
    ((77 - 1) * 10.2^2) + ((15 - 1) * 11.3^2)
    ) /
    (77 + 15 - 2)
  )
s
```

```
## [1] 10.37877
```

Having obtained our pooled s, we can use that in calculating se_{diff}:

$$se_{diff} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 10.37877\sqrt{\frac{1}{77} + \frac{1}{15}}$$

```
se_diff <- 10.37877 * sqrt((1/77) + (1/15))
se_diff
```

```
## [1] 2.929199
```

When using pooled s, our degrees of freedom formula is:

$$df = n_1 + n_2 - 2$$

So our degrees of freedom is:

$$df = 77 + 15 - 2 = 90$$

(f) Perform a significance test for the difference of means under the assumption of equal variances (α = .05).

i) assumptions: Equal variance. Sample sizes are small, so assuming a normal distribution. Means so using a t statistic.
ii) hypothesis: $H_0 : \mu_2 - \mu_1 = 0$, $H_a : \mu_2 - \mu_1 \neq 0$.
iii) Critical value and test statistic: We have $n_1 + n_2 - 2 = 90$ degrees of freedom, so the critical value of t (using 80 df from the t-statistic table) for $\alpha = .05$ is: $t_{.025} = 1.990$. To calculate the test statistic:

$$t = \frac{(\bar{x_2} - \bar{x_1}) - 0)}{se_{diff}} = \frac{71.2 - 64.8 - 0}{2.929199} = 2.184898$$

iv) p-value: Finding the closest t value at 80 degrees of freedom, and multiplying the confidence level ranges by 2: $t_{.020} < p < t_{.050}$.
v) conclude: Since the absolute value of observed t-statistic exceeds the critical value (1.990), and similarly, the p-value is less than α, we can reject the null hypothesis that the difference between the two means is zero.

3. In the `anes2020subset` dataset, the variable `BAplus` is a dichotomous variable in which 1 indicates the person has a BA degree or higher, and 0 indicates the person does not. In this question, you will test whether the mean of `PoliceTherm` is different for the two populations represented by these samples. `PoliceTherm` is a person's feeling thermometer score for the police.

```
# Note: code for 3.a:
aggregate(x = PoliceTherm ~ BAplus,
          data = anes,
          FUN = function(x) c("mean" = mean(x),
                              "SD" = sd(x),
                              "freq" = length(x)))
```

(a) Make a table that shows the means of `PoliceTherm` for each category of `BAplus`. In Stata, the command is `tab BAplus, summarize(PoliceTherm)`. In R, the most simple command is `aggregate(PoliceTherm ~ BAplus, data = anes2020, FUN = mean)`. This does not give you standard deviations or frequencies, however. So, I encourage you to go to page 10 of the R help document on Canvas and use the example in the middle of the page as the basis for your command.

```
##   BAplus PoliceTherm.mean PoliceTherm.SD PoliceTherm.freq
## 1      0         72.29115       26.07943       3943.00000
## 2      1         68.60240       23.79355       3335.00000
```

```
# note: code for question 3.b:
t.test(anes$PoliceTherm ~ anes$BAplus)
```

(b) Use software to perform a significance test for the difference of means. In Stata, the command in Stata is `ttest PoliceTherm, by(BAplus) unequal`. In R, the command is `t.test(anes2020$PoliceTherm ~ anes2020$BAplus)`

```
## 
##  Welch Two Sample t-test
## 
## data:  anes$PoliceTherm by anes$BAplus
## t = 6.3054, df = 7234.4, p-value = 3.044e-10
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  2.541942 4.835558
## sample estimates:
## mean in group 0 mean in group 1 
##        72.29115        68.60240
```

(c) Report the difference between the means and state the conclusion of this significance test. The difference in means is -3.68875. The test returns a 95% confidence interval that the true difference in means is between 2.541942 and 4.835558. The p-value of 0.0000000003044 is much smaller than .05, suggesting that we should reject the null hypothesis that the true difference between the two means is zero.

4. Continuing with the `anes2020subset` dataset, this question will have you test whether the proportion of people who have health insurance is different for people who have a BA degree versus those who do not have a BA degree.

```
# note: this is code for question 4.a:
freq_table <- table(anes$HealthIns, anes$BAplus)
rownames(freq_table) <- c("Does not have Health Ins.", "Has Health Ins.")
colnames(freq_table) <- c("Does not have a BA", "Has BA")
freq_table <- addmargins(freq_table)
freq_table
```

(a) Make a joint frequency distribution table with row and column totals for these two variables. Have `HealthIns` make the rows and `BAplus` make the columns. You have made this kind of table before.

```
##
##                                 Does not have a BA Has BA  Sum
##   Does not have Health Ins.                    607    202  809
##   Has Health Ins.                             3840   3409 7249
##   Sum                                         4447   3611 8058
```

(b) Do this question by hand using the formulas. Using the data from the table, perform a significance test for whether the proportion of people that have health insurance is different across the categories of `BAplus`. Be sure to use the correct formula for the standard error.
This is a comparison of two proportions. First we need to calculate our proportions from the “Has Health Insurance” row of the frequency table for each category of the `BAplus` variable:

$$\text{(Has Health Ins. and No BA)} = \frac{3840}{4447} = 0.8635035$$
$$\text{(Has Health Ins. and BA)} \;= \frac{3409}{3611} = 0.9440598$$

i) assumptions: Independent samples, with normal distribution. Two proportions so using a z test. No confidence level is given so assume using $\alpha = .05$
ii) hypothesis: $H_0 : \hat{\pi}_1 - \hat{\pi}_2 = 0$. $H_a : \hat{\pi}_1 - \hat{\pi}_2 \neq 0$
iii) Critical value and test statistic:

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{se_0}, \text{where } se_0 = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}$$

$\hat{\pi}$ can be obtained from the table as the overall proportion that has health insurance:

```r
7249 / 8058
```

```
## [1] 0.8996029
```

$$se_0 = \sqrt{\frac{0.8996029(1-0.8996029)}{4447} + \frac{0.8996029(1-0.8996029)}{3611}}$$

```r
se_0 <-sqrt(
  ((0.8996029 * (1 - 0.8996029)) / 4447) +
    ((0.8996029 * (1 - 0.8996029)) / 3611)
  )
se_0
```

```
## [1] 0.006732127
```

Using se_0 in the z score calculation:

$$z = \frac{0.8635035 - 0.9440598}{0.006732127} = -11.96595$$

iv) p-value: Our observed test statistic is -11.96595, so we know that our p-value is less than .000000287*2 = 0.000000574 v) conclude: Since the absolute value of our observed test statistic exceeds the critical value (1.96), we can reject the null hypothesis that the difference of the proportions is zero. Similarly, since the p-value is less than .05, we can reject H_0.

```
# note: code for question 4.c:
# x = For Yes Health Ins row: No BA & BA
# n = Total No BA, Total BA
mytest <- prop.test(x = c(3840, 3409),
          n = c(4447, 3611),
          correct = FALSE)
mytest
```

(c) Now, use software to perform this test. In Stata, the command in Stata is `prtest HealthIns, by(BAplus)`. In R, the process is more involved. Follow the steps outlined in Section 8.2 of the R Help document.

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(3840, 3409) out of c(4447, 3611)
## X-squared = 143.18, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.09312602 -0.06798664
## sample estimates:
##    prop 1    prop 2
## 0.8635035 0.9440598
```

In the output, `prop.1` is the proportion of people without a BA who report that they have health insurance. `prop.2` is the proportion who have health insurance with a BA. The χ^2 statistic is 143.18, and the p-value is less than 0.00000000000000022. Thus we can reject the null hypothesis that the difference between the proportions is zero. To obtain the z-statistic:

```
z_stat <- sqrt(mytest$statistic)
names(z_stat) <- "z-statistic"
z_stat
```

```
## z-statistic
##    11.96595
```

Which matches our manually calculated value!!!!!!!

5. Use the dataset `LifeExpectancy` for this question. In this dataset, the variable `LifeExp2000` is the level of life expectancy for each country in the year 2000. The variable `LifeExp2010` is the level of life expectancy for the same set of countries in the year 2010. These are dependent samples.

```
# note: code for question 5.a:
life <- read_dta('LifeExpectancy.dta')
life$LEDiff <- life$LifeExp2010 - life$LifeExp2000
mean(life$LEDiff)
```

(a) To measure the difference in these two life expectancy rates for each country, use the generate command to create a new variable called `LEdiff`. In Stata, use the generate command: `gen LEdiff = LifeExp2010 - LifeExp2000`. In R, the command is: `life$LEdiff <- life$LifeExp2010 - life$LifeExp2000`. Browse your results with either browse in Stata or `View(life)` in R. When done, use commands to obtain the mean, standard deviation, and sample size for `LEdiff`. Report your findings in your answers. How do you interpret the mean of `LEdiff`?

```
## [1] 2.92028
```

```
sd(life$LEDiff)
```

```
## [1] 1.805962
```

```
length(life$LEDiff)
```

```
## [1] 181
```

The mean difference in life expectancy in 2010 vs. 2000 is 2.92028 years. The standard deviation is 1.805962 across the sample. There are 181 records in the data set.

```
# note: code for question 5.b:
mean2000 <- mean(life$LifeExp2000)
mean2010 <- mean(life$LifeExp2010)
mean2010 - mean2000
```

(b) We learned that the mean of the differences is the same as the difference of the means. Is that true? Use commands to find the means of `LifeExp2010` and `LifeExp2000`, then calculate the difference between them. Compare this to the mean of `LEdiff` that you found above.

```
## [1] 2.92028
```

The subtracting `mean2000` from `mean2010` results in 2.92028. This is the same as the `LEdiff` value found above.

(c) This insight from part (b) tells us that testing whether the mean of `LEdiff`=0 is the same as testing whether `LifeExp2010` and `LifeExp2000` have different means. By hand, perform the test of whether the mean of `LEdiff`=0. You have the information you need to calculate the standard error from your summary statistics in part (a). It's a one-sample test of statistical significance. Produce a t statistic and p-value for this test.

i) assumptions: large random sample, normal distribution. Means so using t statistic.
ii) hypothesis: $H_0 : \mu = 0$, $H_a : \mu \neq 0$.
iii) critical value and test statistic: At $\alpha = .05$, using df = 100 (180 is not on the table), the critical value of t is 1.984. To calculate our test statistic:

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}, \text{ where } \hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{1.805962}{\sqrt{181}} = 0.1342361$$

$$t = \frac{2.92028 - 0}{0.1342361} = 21.7548$$

```r
t_stat <- (2.92028 - 0) / (0.1342361)
t_stat
```

```
## [1] 21.7548
```

vi) p-value: Using the t-table and our degrees of freedom, the closest t value we can find is 3.174. Since our observed t-statistic is 21.7548: $ p < t_{.002}$

vii) conclude: Since the absolute value of our observed t-statistic exceeds the critical value, we can reject the null hypothesis that the true mean (difference) is zero. Similarly, our p-value is less than α at the .05 confidence level.

```r
# note: this is code for question 5.d:
t.test(life$LifeExp2000, life$LifeExp2010, paired = TRUE)
```

(d) Now use your software to perform a dependent samples (i.e. paired) t-test for the difference of means. Report the results and compare them to the test that you performed in part (c). In Stata, the command is: `ttest LifeExp2010=LifeExp2000`. In R, the command is: `t.test(life$LifeExp2000, life$LifeExp2010, paired = TRUE)`

```
##
##  Paired t-test
##
## data:  life$LifeExp2000 and life$LifeExp2010
## t = -21.755, df = 180, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -3.185159 -2.655401
## sample estimates:
## mean difference
##       -2.92028
```