Problem Set 8

Connor Tragesser

2023-10-22

Discussed with Alice Moon, Ryo Sawayama, SoYun Chang, Jae Eun Jun, Jinglong Guo, and Jae Chang.

Begin by preparing the environment. The working directory needs to be set where your data will be available (meaning, you will need to change this for your computer), and the packages that are necessary for this exercise need to be loaded from the library (this assumes they are already installed). I'm also clearing the environment before beginning.

```r
library(foreign)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(modelsummary)
library(formatR)
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```r
library(readxl)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
knitr::opts_chunk$set(echo = TRUE)
```

```
knitr::opts_knit$set(root.dir = "/Users/connortragesser/Documents/Political Science Coursework Readings/

getwd()
```

```
## [1] "/Users/connortragesser/Documents/quant_1"
```

```
rm(list = ls())

set.seed(2453)
```

Problem 1

Part A

The null hypothesis usually means "the expected value of a parameter if there is no effect or relationship."

Part B

We have dependent samples.

Part C

We use the pooled estimate of the proportion in two samples to test the null hypothesis of no difference.

Part D

The evidence against the null hypothesis grows stronger.

Part E

In a test of hypotheses, the p-value is the probability, assuming the null hypothesis is true, that a test statistic will take value at least as extreme as what is actually observed.

Problem 2

Gailmard 8.1

Part A

I will use a two sample difference in means test. The null hypothesis is that the average PC income is the same between the Obama and McCain states, and the alternative hypothesis is that they will be different. It will be a two-tailed test.

the equation below is used to calculate the degrees of freedom for the t-distribution in this case

$$\frac{(\frac{S_1^2}{N_1}+\frac{S_2^2}{N_2})^2}{(\frac{S_1^2}{N_1})^2/(N_1-1)+(\frac{S_2^2}{N_2})^2/(N_2-1)}$$

```
t <- (44500 - 38500)/(sqrt((5400^2/29) + (4600^2/21)))
dof <- (((5400^2/29) + (4600^2/21))^2)/(((((5400^2/29))^2)/(29 -
    1)) + ((((4600^2/21))^2)/(21 - 1)))

2 * pt(t, df = dof, lower.tail = FALSE)
```

```
## [1] 0.0001086105
```

This t score allows us to reject the null hypothesis that the average PC income is the same between the Obama and McCain states.

Part B

It does not. Andrew Gelman notes this phenomenon in his book Red State Blue State Rich State Poor State, calling it the "red-blue paradox" but it is an application of Simpson's Paradox. In short, Simpson's paradox describes phenomena where aggregation of groups of data mask underlying effects. The proposition that people could still tend to vote republican as they get richer would still hold if the shift is heterogeneous between states. As in, the increase is smaller in blue states and larger in red states.

Gailmard 8.2

Part i

The null hypothesis is that the proportion that comes up heads is .5

$$H_0 : \hat{\pi} = .5$$
$$H_1 : \hat{\pi} \neq .5$$

```r
z <- ((5/14) - 0.5)/sqrt((0.5 * 0.5)/14)

2 * pnorm(z)
```

```
## [1] 0.2850494
```

This test shows that we fail to reject the null hypothesis that the coin is fair.

Part ii

Binomial probability

Calculate the probability of observing this under the null hypothesis of $H_0 : \hat{\pi} = .5$

```r
choose(14, 5) * (0.5^5) * (0.5^9)
```

```
## [1] 0.1221924
```

Thus, we fail to reject the null hypothesis at the $\alpha = 0.05$ level, we do not have enough evidence that our observed π goes against our assumption.

Problem 3

Begin by loading dataset (It should be in the folder you set the WD to at the beginning).

```r
nes <- read.dta("nes2004c.dta")
head(nes)
```

```
##     gender                                     race rep_therm libcon3_r partyid7
## 1   1 Male 50. White (no mention of other race)       100         3        6
## 2   1 Male 50. White (no mention of other race)       100         3        6
## 3 2 Female                                10. Black         0         2        0
## 4   1 Male                                10. Black        85        NA        2
## 5   1 Male                                10. Black        70        NA        1
## 6   1 Male                                10. Black        70        NA        4
##   attend3 gay_marriage south white
## 1       1            0     1     1
## 2       3            0     1     1
## 3       1            0     0     0
## 4       1            0     1     0
## 5       1            0     1     0
## 6       1            0     0     0
```

Part A

```r
nes %>%
    group_by(south) %>%
    summarise(mean(rep_therm, na.rm = TRUE)) %>%
    ungroup()
```

```
## # A tibble: 2 x 2
##   south `mean(rep_therm, na.rm = TRUE)`
##   <dbl>                           <dbl>
## 1     0                            51.9
## 2     1                            55.8
```

```r
t.test(rep_therm ~ south, data = nes)
```

```
##
##  Welch Two Sample t-test
##
## data:  rep_therm by south
## t = -2.3107, df = 732.06, p-value = 0.02113
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -7.3402127 -0.5968076
## sample estimates:
## mean in group 0 mean in group 1
##        51.87726        55.84577
```

We reject the null hypothesis that there is no difference in feelings toward Republicans between Southerners and non-Southerners at the $\alpha = .05$ level, but not at the $\alpha = .01$ level.

Part B

```r
t.test(rep_therm ~ south, data = nes, subset = white == 1)
```

```
##
##  Welch Two Sample t-test
##
## data:  rep_therm by south
## t = -4.4385, df = 437.79, p-value = 1.148e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -13.105905  -5.061359
## sample estimates:
## mean in group 0 mean in group 1
##        52.80661        61.89024
```

```r
t.test(rep_therm ~ south, data = nes, subset = white != 1)
```

```
##
##  Welch Two Sample t-test
##
## data:  rep_therm by south
## t = 0.75954, df = 298.23, p-value = 0.4481
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -3.557745  8.030132
## sample estimates:
## mean in group 0 mean in group 1
##         48.5503         46.3141
```

Among whites, we reject the null hypothesis that there is no difference in feelings toward the Republican party between Southerners and non-Southerners $\alpha = .05$ and $\alpha = .01$ levels. We fail to reject that same null hypothesis among non-whites at the $\alpha = .05$ level

Problem 4

Download dataset:

```
pakistan <- read_xlsx("pakistan.xlsx", sheet = "final")

pakistan$hhincome <- as.numeric(pakistan$hhincome)

pakistan$ln_hhincome = log(pakistan$hhincome)
```

Part A

The null hypothesis, or H_0, is that the average log-household income for Pakistani households is log-38,000 Pakistani Rupees/year. The alternative Hypothesis H_1, is that the average log-household income is different from log-38,000 in either direction. We are calculating the one sample t-test. Put symbolically

$$H_0 : ln(\mu) = ln(38,000)$$
$$H_1 : ln(\mu) \neq ln(38,000)$$

So we need to define a test statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$
$$t = \frac{ln(\bar{x}) - ln(38000)}{s/\sqrt{12}}$$

```
(mean(pakistan$ln_hhincome) - log(38000))/(sd(pakistan$ln_hhincome)/sqrt(12))
```

```
## [1] -2.452667
```

Part B

It is important that we assume that the natural log of income is distributed normally in order to conduct a hypothesis test of the sample mean with a sample size of 12. We do not know the population variance, σ, of income. In that situation, with a small sample, if the DGP was not normal, we could not construct a distribution in order to conduct a hypothesis test. We simply would not have enough information. However, with a small sample, if we assume normality, we can use the t distribution in order to conduct a hypothesis test. Assuming the null hypothesis is true, the test statistic should be a t-distribution centered at log(38,000)

Part C

```
t.test(pakistan$ln_hhincome, mu = log(38000))
```

```
##
##  One Sample t-test
##
## data:  pakistan$ln_hhincome
## t = -2.4527, df = 11, p-value = 0.03209
## alternative hypothesis: true mean is not equal to 10.54534
## 95 percent confidence interval:
##   8.691266 10.445068
## sample estimates:
## mean of x
##  9.568167
```

This test shows that we reject the null hypothesis that the log mean of household income in Pakistan is equal to log(38,000) at the $\alpha = .05$ level.