PSYC2020L Assignment 8

[name]

2025-10-10

Goals for Review:

- Importing
- NHST

Setting Up

1) Rename this assignment as “Lab 8 Assignment [Last, First].Rmd”

2) Check and set (if needed) your working directory. If you’d like to load packages with `library()`, please do so in this block. Make sure to load the `tidyverse` package!

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   4.0.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
here::here()
```

```
## [1] "C:/Users/jhelmer3/OneDrive - Georgia Institute of Technology/Courses/GTA/PSYC2020L"
```

Import

3) Download the `testdata.csv` file from Canvas, and relocate it to the folder for this lab. Import the csv into R. Try knitting the document right after this question for simpler troubleshooting! This data contains three variables:

name	description	notes
`id`	examinee identifier	
`SAT`	SAT score	can theoretically range 400–1600
`ACT`	ACT score	can theoretically range 1–36

Display the first few rows of this data.

```
dat <- rio::import(here::here("labs", "Lab 8 - NHST II", "testdata.csv"))
head(dat)
```

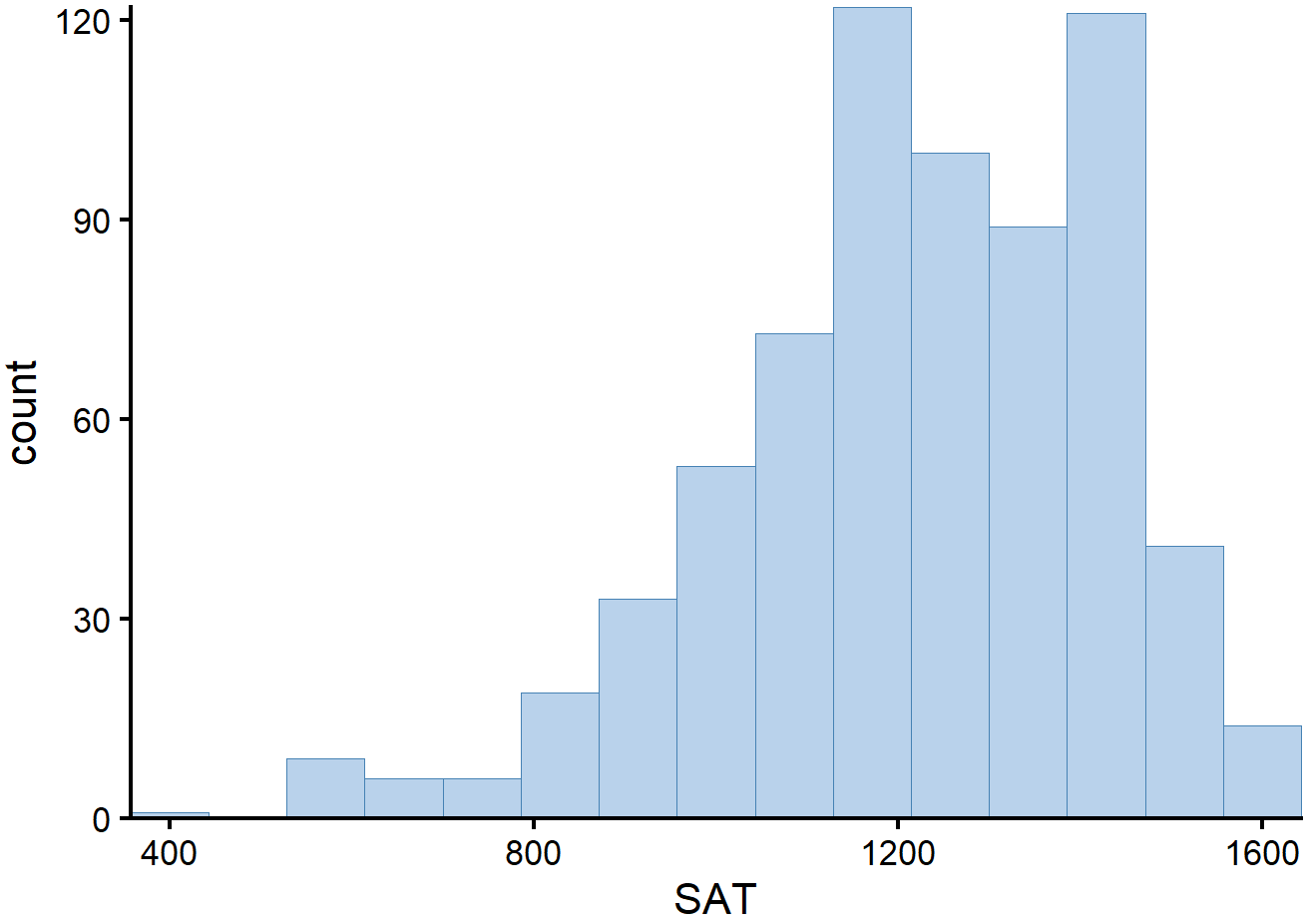
```
##   id  SAT ACT
## 1  1 1000  24
## 2  2 1100  35
## 3  3  950  21
## 4  4 1070  26
## 5  5 1150  31
## 6  6 1280  28
```

EDA

4) Explore or report on the SAT variable in three different ways. Examples include making a plot, commenting on the shape of the distribution, numerical summary statistics (e.g, mean, SD, range), etc.

```
ggplot(dat, aes(x = SAT)) +
  geom_histogram(bins = 15, fill = "slategray2", color = "steelblue", linewidth = .2) +
  coord_cartesian(expand = F) +
  theme_classic(base_size = 16)
```

```
## Warning: Removed 13 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
range(dat$SAT, na.rm = T)
```

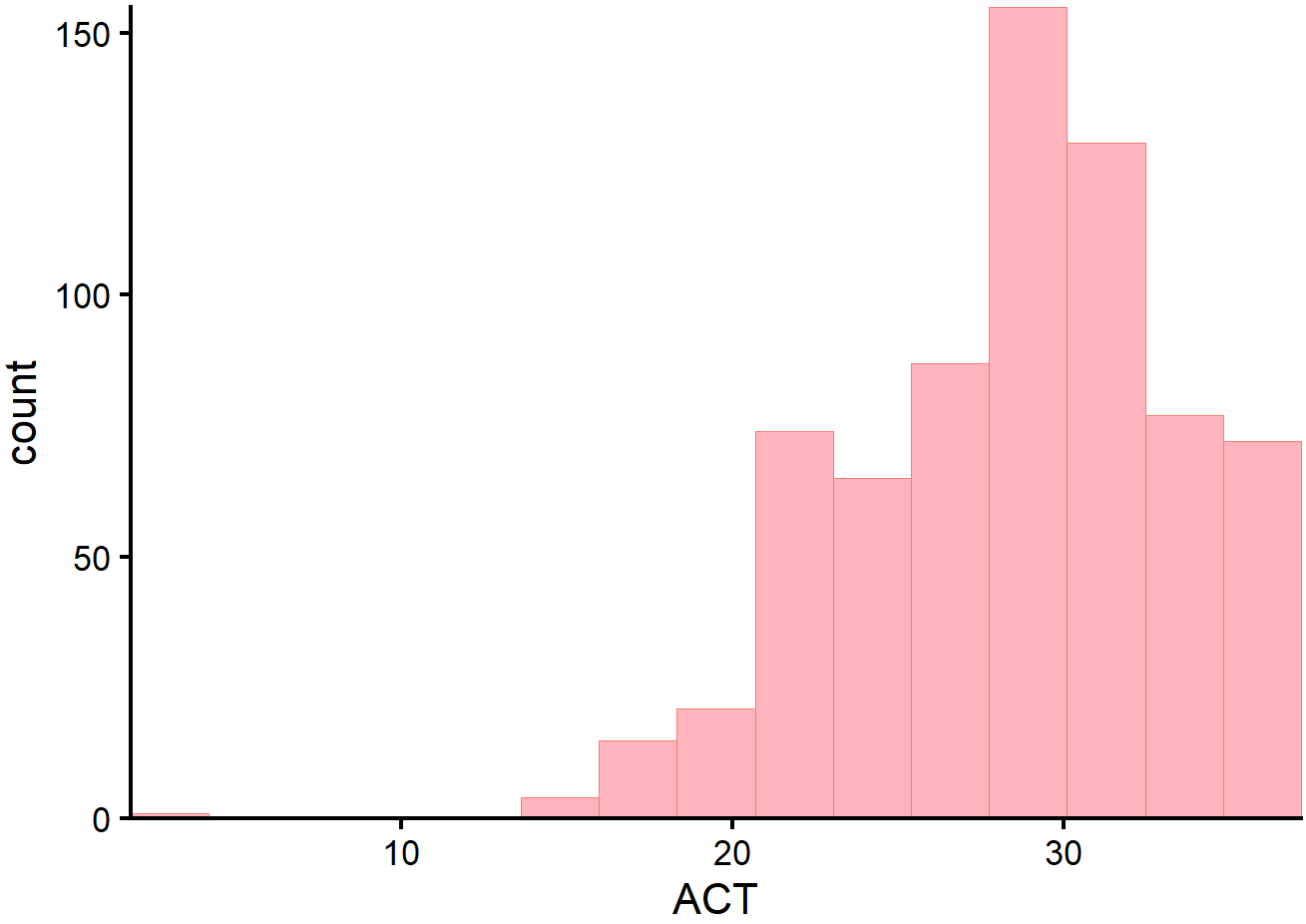
```
## [1]  400 1600
```

```
mean(dat$SAT, na.rm = T)
```

```
## [1] 1222.552
```

4) Explore or report on the ACT variable in three different ways. Examples include making a plot, commenting on the shape of the distribution, numerical summary statistics (e.g, mean, SD, range), etc.

```
ggplot(dat, aes(x = ACT)) +
  geom_histogram(bins = 15, fill = "lightpink", color = "salmon", linewidth = .2) +
  coord_cartesian(expand = F) +
  theme_classic(base_size = 16)
```




```
range(dat$ACT)
```

```
## [1]  3 36
```

```
mean(dat$ACT)
```

```
## [1] 28.54714
```

Z-Tests

You've been tasked with testing whether the scores from this sample of students differ from the population means.

5a) Let's say the population mean for SAT scores is 1200 and the population SD is 110. What are your null and alternative hypotheses?

The null hypothesis is that the observed mean of SAT scores is equal to the hypothesized population mean of 1200. The alternative is that the observed mean of SAT scores is not equal to the hypothesized population mean of 1200.

5b) Calculate the *z*-test statistic for this sample of SAT scores against the population.

```
(mean(dat$SAT, na.rm = T) - 1200) / (110 / sqrt(sum(!is.na(dat$SAT))))
```

```
## [1] 5.373589
```

```
(mean(dat$SAT, na.rm = T) - 1200) / (110 / length(dat$SAT)) # ideally they do the top one but th
is is finee
```

```
## [1] 143.5107
```

5c) What is the *p*-value of that *z*-statistic?

```
pnorm((mean(dat$SAT, na.rm = T) - 1200) / (110 / sqrt(sum(!is.na(dat$SAT)))), lower.tail = F)
```

```
## [1] 3.859228e-08
```

```
pnorm((mean(dat$SAT, na.rm = T) - 1200) / (110 / length(dat$SAT)), lower.tail = F) # for if they
do the top one
```

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

5d) Provide an interpretation for this *z*-test given the test statistic and *p*-value.

We reject the null hypothesis. There is a significant difference between the observed sample mean and the population mean.

6a) Let's say the population mean for ACT scores is 29 and the population variance is 64. What are your null and alternative hypotheses?

The null hypothesis is that the observed mean of ACT scores is equal to the population mean of 28. The alternative is that the observed mean of ACT scores is not equal to the hypothesized population mean of 28.

6b) Calculate the *z*-test statistic for this sample of ACT scores against the population.

```
(mean(dat$ACT, na.rm = T) - 29) / (sqrt(64) / sqrt(sum(!is.na(dat$ACT))))
```

```
## [1] -1.497684
```

```
(mean(dat$ACT, na.rm = T) - 29) / (sqrt(64) / length(dat$ACT)) # is fineeeeee
```

```
## [1] -39.625
```

6c) What is the *p*-value of that *z*-statistic?

```
pnorm((mean(dat$ACT, na.rm = T) - 29) / (sqrt(64) / sqrt(sum(!is.na(dat$ACT)))))
```

```
## [1] 0.06710766
```

```
pnorm((mean(dat$ACT, na.rm = T) - 29) / (sqrt(64) / length(dat$ACT))) # okayyyy if use the lengt
h() function before
```

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

6d) Provide an interpretation for this *z*-test given the test statistic and *p*-value.

We retain the null hypothesis. In this sample we do not see a significant difference between the observed sample mean and the population mean of ACT scores.