Homework #1

Yunting Chiu

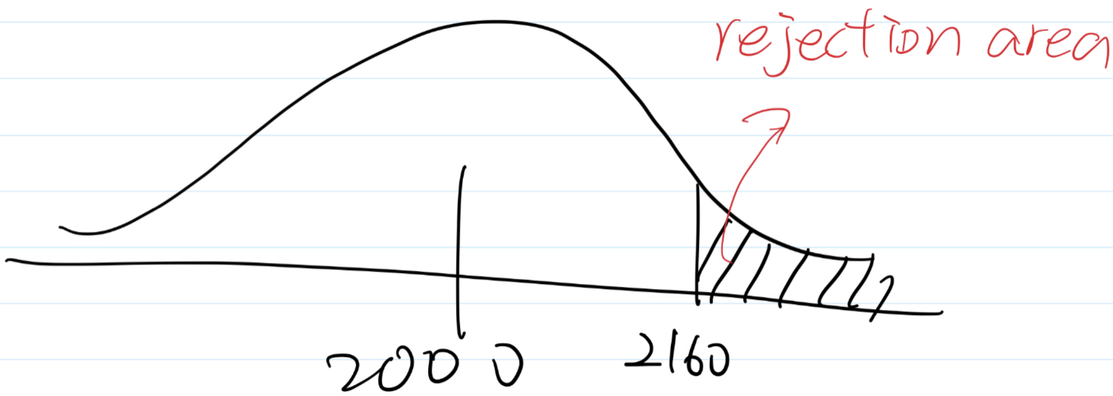
2021-01-24

Review of Estimation and Hypothesis Testing (handouts, your old notes, ...)

When α is not given, use the p-value approach to make your conclusions. When it's difficult to conclude, use $\alpha = 0.05$. For two-sample problems, use the F-test to decide which t-test to use.

1. The manufacturer of a certain brand of household light bulbs claims that the bulbs produced by his factory have an average life of at least 2,000 hours. The mean and standard deviation of 20 light bulbs selected from the manufacturer's production process were calculated to be 2,160 and 142 hours, respectively.

(a) Do the data represent sufficient evidence to support the manufacturer's claim? How can you interpret your answer?



- Let we set H0: mu < 2000, and Ha: mu >= 2000. Note, this is a one-sided t.test with $\alpha = 0.05$, and degree of freedom: n-1 = 20-1 = 19. According to the information given, the sample mean is 2160, and the sample standard deviation is 142. We use t-statistics to conduct the statistical inference.

$$t = \frac{\hat{\mu} - \mu}{s/\sqrt{n}} = \frac{2160 - 2000}{142/\sqrt{20}} = 5.04$$

- The critical value for $\alpha = 0.05$ at df = 19 is 1.729 according to the t-distribution table. As 5.05 > 1.729, so we reject the null and claim that we have sufficient evidence to support the manufacturer's claim with a 5 % probability that I am going to reject the null when it is true.

(b) Construct a 95% confidence interval for the mean lifetime of household light bulbs.

95% confidence interval

$$= \hat{\mu} \pm t(S/\sqrt{n})$$

$$= 2160 \pm 1.729 \cdot \frac{142}{\sqrt{20}}$$

$$= (2105.10\ ,\ 2214.90)$$

2. There are two manufacturing processes, old and new, that produce the same product. The defect rate has been measured for 20 days for the old process, and for 14 days for the new process, resulting in the following sample summaries.

——	**OLD**	**NEW**
Average defect rate	4.7	2.3
Standard deviation	6.8	5.0

The firm is interested in switching to the new process only if it can be demonstrated convincingly that the new process reduces the defect rate. Is there significant evidence of that? Use $\alpha = 5\%$; assume that the collected data represent two random samples from Normal distributions. Use the method of testing that is appropriate for this situation.

3. (Required for Stat-615, optional for Stat-415) An account on server A is more expensive than an account on server B. However, server A is faster. To see if whether it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is. A certain computer algorithm is executed 30 times on server A and 20 times on server B with the following results,

(a) Is there a significant difference between the two servers?
(b) Is server A significantly faster?

4. Micro-project. Data on 522 recent home sales are available on our Blackboard web site The following variables are included.
   Use software to find out if there is significant evidence that:

(a) The sales price depends on the air conditioner in the house.

(b) On the average, homes with an air conditioner are more expensive.

(c) On the average, homes with an air conditioner are larger.

(d) The sales price depends on the proximity to a highway.

(e) On the average, homes are cheaper when they are close to a highway.

(f) On the average, homes are cheaper when they are far from a highway.

```
# read the dataset
library(tidyverse)
```

```
## -- Attaching packages ---- tidyverse 1.3.0 --
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## -- Conflicts ------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
homeSales <- read_csv("./data/HOME_SALES(1).csv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   SALES_PRICE = col_double(),
##   FINISHED_AREA = col_double(),
##   BEDROOMS = col_double(),
##   BATHROOMS = col_double(),
##   GARAGE_SIZE = col_double(),
##   YEAR_BUILT = col_double(),
##   STYLE = col_double(),
##   LOT_SIZE = col_double(),
##   AIR_CONDITIONER = col_character(),
##   POOL = col_character(),
##   QUALITY = col_character(),
##   HIGHWAY = col_character()
## )
```

```
homeSales
```

```
## # A tibble: 522 x 13
##      ID SALES_PRICE FINISHED_AREA BEDROOMS BATHROOMS GARAGE_SIZE YEAR_BUILT
##   <dbl>       <dbl>         <dbl>    <dbl>     <dbl>       <dbl>      <dbl>
## 1     1         360          3032        4         4           2       1972
## 2     2         340          2058        4         2           2       1976
## 3     3         250          1780        4         3           2       1980
## 4     4         206.         1638        4         2           2       1963
## 5     5         276.         2196        4         3           2       1968
## 6     6         248          1966        4         3           5       1972
```

```
## 7     7        230.          2216        3         2           2       1972
## 8     8        150           1597        2         1           1       1955
## 9     9        195           1622        3         2           2       1975
## 10    10       160           1976        3         3           1       1918
## # ... with 512 more rows, and 6 more variables: STYLE <dbl>, LOT_SIZE <dbl>,
## #   AIR_CONDITIONER <chr>, POOL <chr>, QUALITY <chr>, HIGHWAY <chr>
```