Lab 2

Yunting Chiu

2021-01-29

R Lab 2. Review of T-tests and F-tests

```r
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()
```

```r
H <- read_csv("HOME_SALES.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()
## )
```

```r
head(H)
```

```
## # A tibble: 6 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
## # ... with 6 more variables: STYLE <dbl>, LOT_SIZE <dbl>,
## #   AIR_CONDITIONER <chr>, POOL <chr>, QUALITY <chr>, HIGHWAY <chr>
```

1. A one-sample T-test

1a. A one-sample, two-sided T-test

There is a claim that the average price of homes in the region is $300,000. Does the data set support or disprove the claim? This is a two-sided test because there is no specified direction, we are just testing if the population mean is 300,000 or not.
- With the small p-value 0.0002759, we have sufficient evidence reject the null (H0 is rejected). That is, the mean of home price is not 300,000. The sample of mean is 277,8941 and the mean price is 266,035 - 289,754 dollars with 95 % CI.

```
t.test(H$SALES_PRICE, mu=300)
```

```
##
##  One Sample t-test
##
## data:  H$SALES_PRICE
## t = -3.6619, df = 521, p-value = 0.0002759
## alternative hypothesis: true mean is not equal to 300
## 95 percent confidence interval:
##  266.0348 289.7535
## sample estimates:
## mean of x
##  277.8941
```

```
# compute the t-statistic by hand
n <- length(H$SALES_PRICE)
n
```

```
## [1] 522
```

```
mean(H$SALES_PRICE)
```

```
## [1] 277.8941
```

```
sd(H$SALES_PRICE)
```

```
## [1] 137.9234
```

```
# use the formula
t <-  (mean(H$SALES_PRICE) - 300) / (sd(H$SALES_PRICE)/sqrt(length(H$SALES_PRICE)))
t
```

```
## [1] -3.661884
```

1b. A one-sample, left-tail T-test.

Is the mean price less than $300,000? This is a one-sided, left-tail test.
- We have significant evidence that the mean of home price is less than 300 thousand dollars.

```
attach(H)
t.test(SALES_PRICE, mu=300, alternative="less")
```

```
##
##  One Sample t-test
##
```

```
## data:  SALES_PRICE
## t = -3.6619, df = 521, p-value = 0.000138
## alternative hypothesis: true mean is less than 300
## 95 percent confidence interval:
##      -Inf 287.8414
## sample estimates:
## mean of x
##  277.8941
```

1c. A one-sample, right-tail T-test.

Is there any evidence that the mean price is above $300,000?
- With the large p-value, we fail to reject the null. That is, no evidence conclude that the mean of home price is greater than 300 thousand dollars.

```r
t.test(SALES_PRICE, mu=300, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  SALES_PRICE
## t = -3.6619, df = 521, p-value = 0.9999
## alternative hypothesis: true mean is greater than 300
## 95 percent confidence interval:
##  267.9469      Inf
## sample estimates:
## mean of x
##  277.8941
```

2. A two-sample T-test

Does the sales price depend on the presence of a pool? To answer this question, we have to compare homes with the pool and without it. This is a comparison of two populations, so it is a two-sample test.
- With a small p-value 0.001408 we have sufficient evidence to reject the null in favor of accepting Ha. In other words, the mean prices are different in the population depending on a pool. With 95 % confidence interval, the sample mean of pool house is 352,120 dollars and the ample mean of non-pool house is 272,396 dollars, respectively.

```r
t.test(x=SALES_PRICE[POOL=="YES"], y=SALES_PRICE[POOL=="NO"])
```

```
##
##  Welch Two Sample t-test
##
## data:  SALES_PRICE[POOL == "YES"] and SALES_PRICE[POOL == "NO"]
## t = 3.428, df = 40.546, p-value = 0.001408
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   32.74042 126.70831
## sample estimates:
## mean of x mean of y
##  352.1203  272.3959
```

- The small p-value explains that the home price with the pool is more expensive with 95 % CI.

```r
t.test(x=SALES_PRICE[POOL=="YES"], y=SALES_PRICE[POOL=="NO"], alternative="greater")
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  SALES_PRICE[POOL == "YES"] and SALES_PRICE[POOL == "NO"]
## t = 3.428, df = 40.546, p-value = 0.0007039
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  40.57595      Inf
## sample estimates:
## mean of x mean of y
##  352.1203  272.3959
```

3. A two-sample F-test of variances

This F-test is used to compare variances of two samples and in particular, to decide which two-sample T- test is appropriate – a test that assumes equal variances or the Satterthwaite approximation.
- Firstly, the ratio is 0.968, which is close to 1, meaning that there is no evidence of different variances. Secondly, the p-value is at a significant level, so the T-test of the equal-variances is justified.

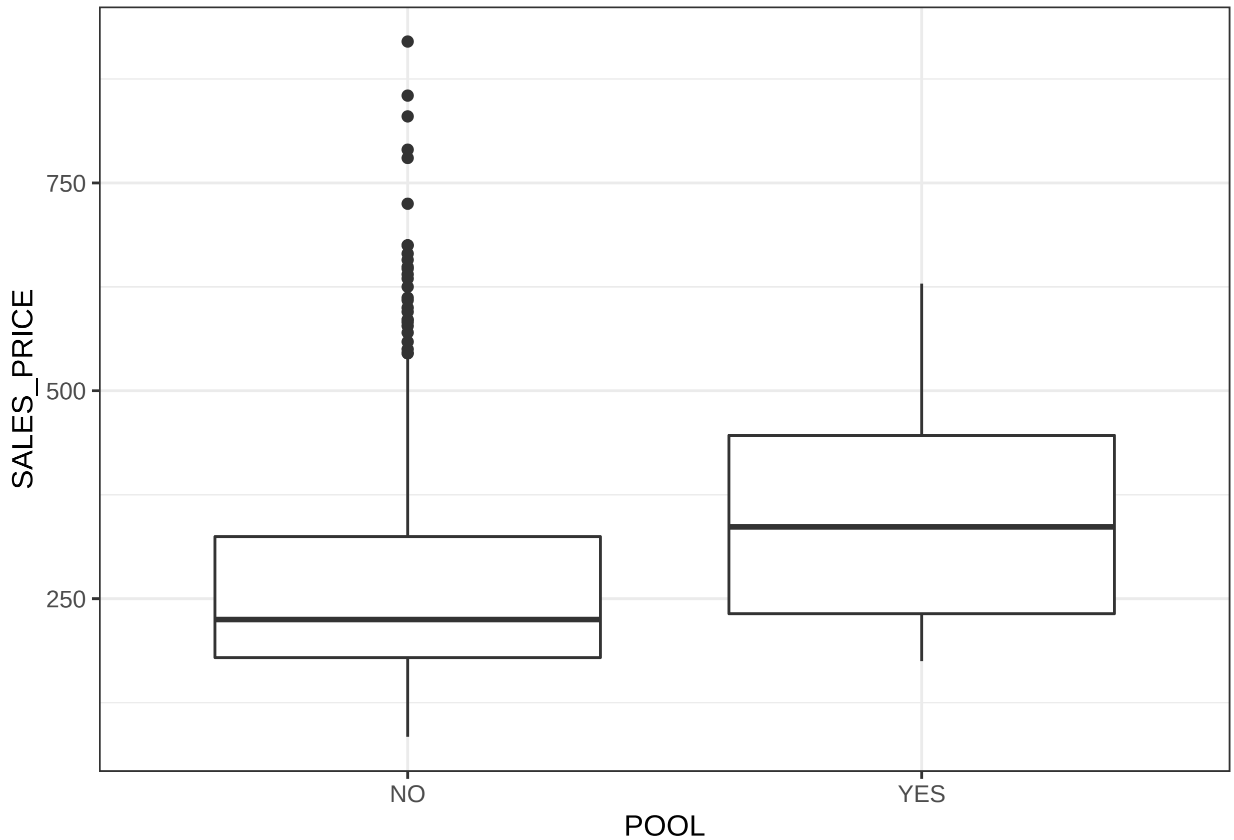
```
var.test(x=SALES_PRICE[POOL=="YES"], y=SALES_PRICE[POOL=="NO"])
```

```
##
##  F test to compare two variances
##
## data:  SALES_PRICE[POOL == "YES"] and SALES_PRICE[POOL == "NO"]
## F = 0.96772, num df = 35, denom df = 485, p-value = 0.9521
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6236526 1.6674409
## sample estimates:
## ratio of variances
##          0.9677224
```

4. Parallel boxplots

- The differences between the two samples can be visualized by parallel box plots. The plot supports our findings on the means and variances.

```
H %>%
  ggplot(aes(x = POOL, y = SALES_PRICE)) +
  geom_boxplot() +
  theme_bw()
```



750
500
250
SALES_PRICE
NO
YES
POOL