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

Set a working directory. Yours is different from mine. It's where you saved the data file from Blackboard.

```
> H = read.csv("HOME_SALES.csv")
> attach(H)
> head(H)
  ID SALES_PRICE FINISHED_AREA BEDROOMS BATHROOMS GARAGE_SIZE YEAR_BUILT STYLE
1  1       360.0          3032        4         4           2       1972     1
2  2       340.0          2058        4         2           2       1976     1
3  3       250.0          1780        4         3           2       1980     1
4  4       205.5          1638        4         2           2       1963     1
5  5       275.5          2196        4         3           2       1968     3
6  6       248.0          1966        4         3           5       1972     1
  LOT_SIZE AIR_CONDITIONER POOL QUALITY HIGHWAY
1    22221             YES   NO  MEDIUM      NO
2    22912             YES   NO  MEDIUM      NO
3    21345             YES   NO  MEDIUM      NO
4    17342             YES   NO  MEDIUM      NO
5    21786             YES   NO  MEDIUM      NO
6    18902             YES  YES  MEDIUM      NO
```

No need to print all 522 rows of data. To get an idea, "head" is a good command, showing the first few lines only.

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.

```
> t.test(SALES_PRICE, mu=300)

        One Sample t-test

data:  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
```

Conclusion: the p-value is very low, hence, there is significant evidence that the mean home price is not $300,000. We also find that the sample mean price in the data set is $277,894, the observed t-statistic is t = -3.66, and the 95% confidence interval for the mean price is [$266,035, $289,754]. You may recall the duality between hypothesis testing and confidence estimation: the level α two-sided test rejects the null hypothesis if and only if the (1- α)100% confidence interval does not contain the tested parameter value. Here we see that the confidence interval does not contain \$300,000, and no surprise, H_0 is rejected.

To compute the t-statistic by hand, we calculated the sample mean and standard deviation
> mean(SALES_PRICE)
```
[1] 277.8941
```
> sd(SALES_PRICE)
```
[1] 137.9234
```

and used the formula for the Student's t-ratio. Or, all in one step,
> t = (mean(SALES_PRICE) - 300) / (sd(SALES_PRICE)/sqrt(length(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 <u>one-sided, left-tail test.</u>

> 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
```

We noticed and explained in class why this p-value is exactly a half of the two-sided p-value. It's very small, so we conclude that yes, there is a significant evidence that the mean home price is less than \$300,000.

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

Is there any evidence that the mean price is *above* \$300,000? Now, this is a <u>one-sided, right-tail test.</u>

> 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
```

This p-value is a complement of the previous one, and it is very high. No evidence that the population mean exceeds $300,000. Certainly! If the sample mean is below 300, it has no way to support a claim that the population mean is above 300.

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.

> 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
```

This test compared the mean of sample X with the mean of sample Y, homes with the pool and without the pool. We find a significant evidence that the mean prices are different in the population, and thus, the price does depend on a pool. The difference between mean prices with and without a pool has 95% confidence limits $32,730 and $126,708. Notice a non-integer number of degrees of freedom. It is calculated by the Satterthwaite approximation.

> 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
```

There is significant evidence that homes with the pool are *more expensive*, on the average.

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.

> 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
```

The ratio of variances is close to 1, and the p-value is high. So, we conclude that there is no evidence of different variances. Thus, the equal-variances T-test is justified.

4. **Parallel boxplots**

We can visualize the differences between the two samples by *parallel boxplots.* When we create a scatterplot with the first variable being categorical, R produces the following. The plot supports our findings about the means and variances.

> plot(POOL, SALES_PRICE)