Day 3

Type I & II Error

Type I: **Rejecting** the null hypothesis when it is **true**

Type II: **Accepting** the null when it is **false**

[**Null Hypothesis is a statement about Population**]

Type I error is called Level of Significance and represented by α

No error situations:

- Accepting the null hypothesis when it true
- Rejecting the null hypothesis when it is false.

Hypothesis Testing

First import files: cs2m, grades, salescity

Next: **install.packages("psych")**

- For comparison: t-test, ANOVA
- If comparison is between two variables/categories then t-test (one sample test, paired sample test, independent sample t-test)
- If comparison is among more than two variables/categories then ANOVA
- To test the association between two variables (categorical variables), Chi-square test

- Golden Rule
  - If the p-value of the test statistics is less than 0.05 (5% level of significance value), reject the null hypothesis.
  - If the p-value of the test statistics is greater than 0.05 (5% level of significance value), fail to reject the null hypothesis.

One Sample t-test

- Null Hypothesis:
  - The null hypothesis assumes that the difference between the true mean (μ) and the comparison value (mo) is equal to zero.

OR

Ho: Average age of the group= 40

One Sample t test

File: cs2m.csv [Ho: Mean Age = 40]

```
> t.test(cs2m$Age, mu=40)
```

```
        One Sample t-test

data:  cs2m$Age
t = -0.6508, df = 29, p-value = 0.5203
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
 30.74814 44.78520
sample estimates:
mean of x
 37.76667
```

Practice (Datafile: Grades)

1. Ho: Average gpa is equal to 1.5.

2. Ho: Average score in quiz 1 is equal to 5.

Paired Sample t-test

- Null Hypothesis
  - The null hypothesis (H0) assumes that the true mean difference (μd) is equal to zero.

OR

$H0: \mu d = 0$

OR

$H0: \mu a = \mu b$

Paired Sample t test

File: grades.csv [Ho: Mean Quiz1 – Mean Quiz 2 = 0]

```
> t.test(x=grades$quiz1, y=grades$quiz2, alternative = "two.sided",
  mu=0, paired = TRUE)
```

```
        Paired t-test

data:  grades$quiz1 and grades$quiz2
t = -2.8717, df = 104, p-value = 0.004948
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8694223 -0.1591491
sample estimates:
mean of the differences
             -0.5142857
```

if p <,= 0.05 (alpha), REJECT

if p> 0.05(alpha), ACCEPT

Paired Sample t test

File: grades.csv [Ho: Mean Quiz1 – Mean Quiz 2 = 0]

```
> t.test(grades$quiz1, grades$quiz2, paired = T)
```

```
        Paired t-test

data:  grades$quiz1 and grades$quiz2
t = -2.8717, df = 104, p-value = 0.004948
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8694223 -0.1591491
sample estimates:
mean of the differences
             -0.5142857
```

Practice

- Ho: The true mean difference between quiz 4 and 5 is equal to zero.

Independent Sample t-test (assuming unequal variance)

- Null hypothesis

  H_0: $\mu_1 = \mu_2$ ("the two-population means are equal")

  GPA (cont.)

  Gender (Male and Female)

  Null: Average GPA of Male = Average GPA of Female

Independent Samples t test {assuming unequal variance}

File: cs2m.csv [Ho: Mean BP across Anxietv Levels are samel

```
> t.test(cs2m$BP~cs2m$AnxtyLH)
```

```
        Welch Two Sample t-test

data:  cs2m$BP by cs2m$AnxtyLH
t = -2.6729, df = 26.613, p-value = 0.01268
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -36.073732  -4.729839
sample estimates:
mean in group 0 mean in group 1
       117.8125        138.2143
```

Independent Samples t test {assuming equal variance}

File: cs2m.csv [Ho: Mean BP across Anxiety Levels are same]

```
> t.test(cs2m$BP~cs2m$AnxtyLH, var.equal=TRUE)
```

```
        Two Sample t-test

data:  cs2m$BP by cs2m$AnxtyLH
t = -2.6897, df = 28, p-value = 0.01192
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -35.93942  -4.86415
sample estimates:
mean in group 0 mean in group 1
       117.8125        138.2143
```

Practice (File: Grades)

- Gender (Categorical)
  - Male (2)
  - Female (1)
- Total Score (Continuous)

Null: Average Total Score of Male = Total Score of Female)

ANOVA
File: salescity.csv

- Null Hypothesis
  - The mean (average value of the dependent variable) is the same for all groups. The alternative or research hypothesis is that the average is not the same for all groups.

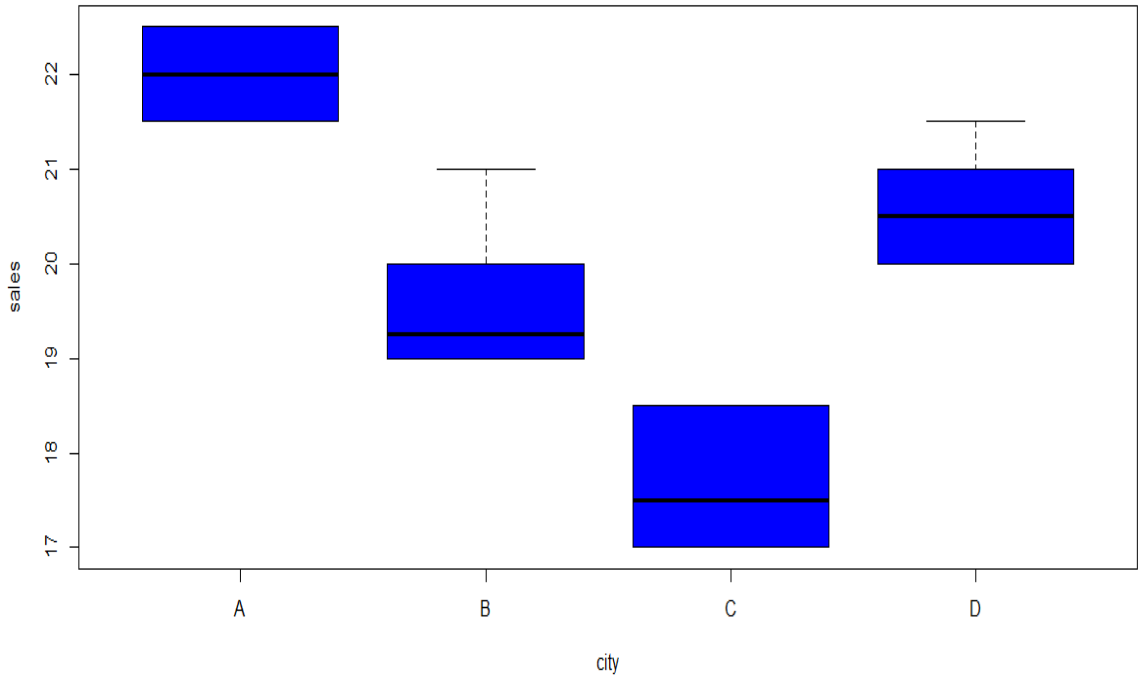
Null hypothesis:
Mean(a)=Mean(b)=Mean(c)=Mean(d)

One Way ANOVA
File: salescity.csv

- Null hypothesis
  - The mean (average value of the dependent variable) is the same for all groups. The alternative or research hypothesis is that the average is not the same for all groups.

Plot Box plots: Sales vs Cities

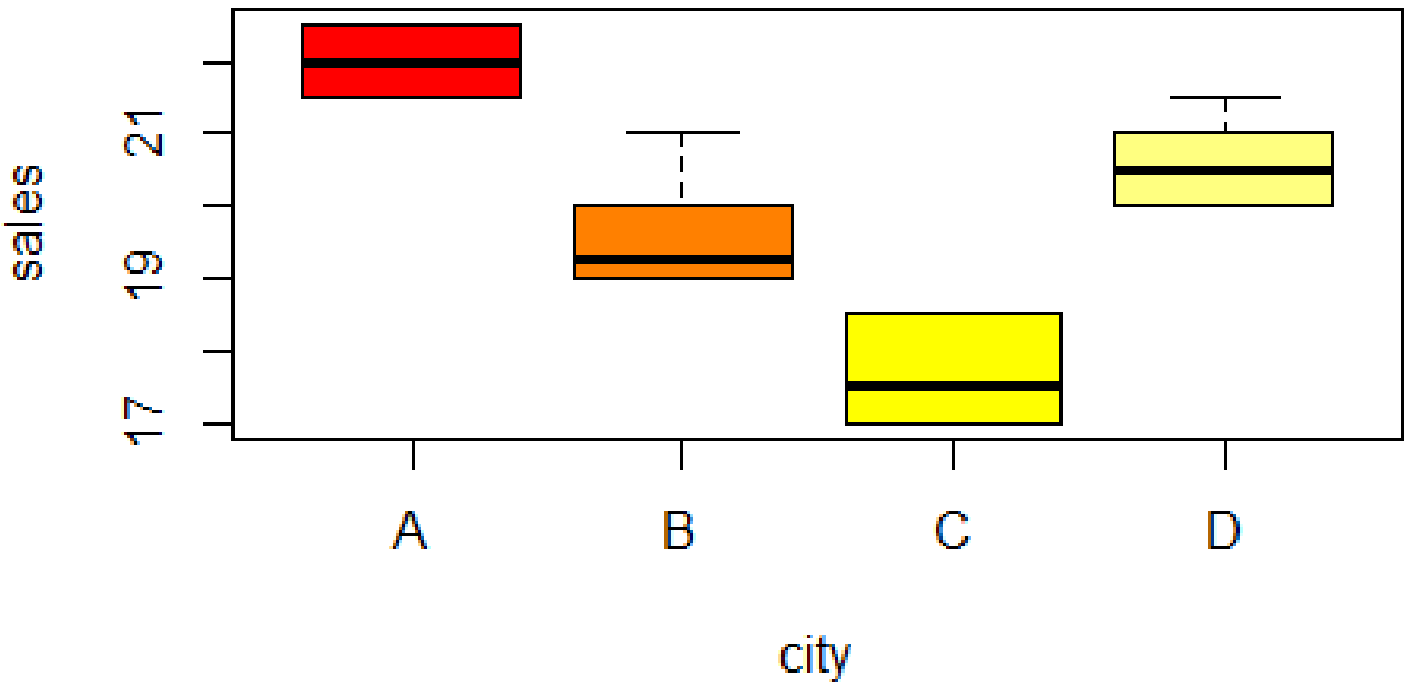
```
> plot(sales~city, data=salescity, col = "blue")
```



```
> plot(sales~city, data = salescity,
 col = heat.colors(4))
```

Ho: Mean sales across cities is same

Ha: at least one city's sale is different from others



ANOVA

```
> results<-aov(sales~city, data = salescity)
> summary(results)
            Df Sum Sq Mean Sq F value   Pr(>F)
city         3  59.71  19.903   43.03 6.54e-09 ***
Residuals   20   9.25   0.462
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> results=aov(sale~city, data=sc)
> summary(results)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
city         3  59.71  19.903   43.03 6.54e-09 ***
Residuals   20   9.25   0.462
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

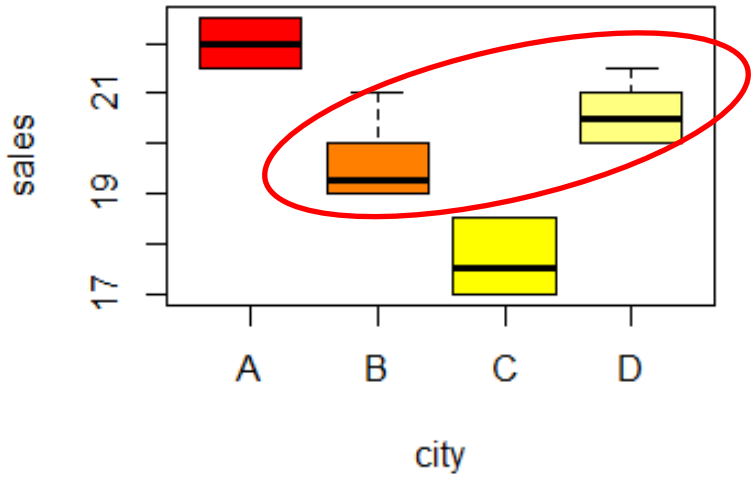
```
> TukeyHSD(results)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = sales ~ city, data = salescity)

$city
         diff         lwr        upr     p adj
B-A -2.416667 -3.51564273 -1.3176906 0.0000286
C-A -4.333333 -5.43230939 -3.2343573 0.0000000
D-A -1.416667 -2.51564273 -0.3176906 0.0087518
C-B -1.916667 -3.01564273 -0.8176906 0.0004849
D-B  1.000000 -0.09897606  2.0989761 0.0826671
D-C  2.916667  1.81769061  4.0156427 0.0000020
```

Ho:
The mean difference across Groups D &B is insignificant

Post Hoc Comparison



Chi-square test

- Null hypothesis
  - No association exists on the categorical variables in the population; they are independent.

Chi-Square Test

File: cs2m [Ho: There is no association between Anxiety and Drug Reaction

```
> chisq.test(cs2m$AnxtyLH, cs2m$DrugR)
```

```
        Pearson's Chi-squared test with Yates' continuity
        correction

data:  cs2m$AnxtyLH and cs2m$DrugR
X-squared = 3.3482, df = 1, p-value = 0.06728
```

Practice

- Ho: No association exists between the gender (Male and Female) and pass/fail status of students (P, F, O).