# Week 10

Analysis of Variance

(ANOVA)

# Recap on week 9

- Paired T-Test
- 2 samples proportion
- Chi-Square test

# Analysis of Variance (ANOVA)

- Levene’s Test (Test for equal variance)
- ANOVA
  - One-way ANOVA
  - Two-way ANOVA
- Post-Hoc test

# One-Way ANOVA

- It is commonly uses to evaluate differences between 3 or more samples.
- The hypothesis will be stated by the difference between the population.
- The null and alternative hypothesis of ANOVA test is stated as the following statement

$H_0$: There are no differences between the population

$H_a$: At least one population is difference

# One-Way ANOVA

- There are 3 assumptions of ANOVA
  - Data are normally distributed
  - Data are independently
  - Data have equal variance (Levene’s Test)

# Levene's Test

- It's the statistical technique to check that variances are equal for all samples.
- Normally, the null and alternative hypothesis can be stated as the following

$H_0$: All population variances are equally

$H_a$: At least one population variance is difference

# One-Way ANOVA example

Food Delivery: In order to compare the average time delivery of three different periods of time (Morning, Lunch and Dinner). The researcher randomly select 10 delivery services of each periods of time and the data are provided as the following:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Breakfast | 15 | 12 | 13 | 14 | 18 | 17 | 17 | 15 | 19 | 25 |
| Lunch | 30 | 32 | 35 | 33 | 34 | 30 | 40 | 28 | 39 | 35 |
| Dinner | 42 | 44 | 43 | 39 | 38 | 45 | 39 | 47 | 43 | 35 |

# Levene’s Test in Python

```
import scipy.stats as stats
group1 = [15, 12, ⋯, 19, 25]
group2 = [30, 32, ⋯, 39, 35]
group3 = [42, 44, ⋯, 43, 35]

levene, pvalue = stats.levene(group1, group2, group3, center = ‘median’

print(“The levene value is “, levene)
print(“The p-value is “, pvalue)
```

# One-way ANOVA table

**Analysis of Variance(ANOVA)**

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares (MS) | F |
|---|---|---|---|---|
| Within | $SS_w = \sum_{j=1}^{k}\sum_{j=1}^{l}(X-\bar{X}_j)^2$ | $df_w = \mathrm{k}-1$ | $MS_w = \frac{SS_w}{df_w}$ | $F = \frac{MS_b}{MS_w}$ |
| Between | $SS_b = \sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2$ | $df_b = \mathrm{n}-\mathrm{k}$ | $MS_b = \frac{SS_b}{df_b}$ | |
| Total | $SS_t = \sum_{j=1}^{n}(\bar{X}_j - \bar{X})^2$ | $df_t = \mathrm{n}-1$ | | |

# One-Way ANOVA example

Food Delivery: In order to compare the average time delivery of three different periods of time (Morning, Lunch and Dinner). The researcher randomly select 10 delivery services of each periods of time and the data are provided as the following:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Breakfast | 15 | 12 | 13 | 14 | 18 | 17 | 17 | 15 | 19 | 25 |
| Lunch | 30 | 32 | 35 | 33 | 34 | 30 | 40 | 28 | 39 | 35 |
| Dinner | 42 | 44 | 43 | 39 | 38 | 45 | 39 | 47 | 43 | 35 |

# One-way ANOVA in Python

```
import scipy.stats import f_oneway
group1 = [15, 12, ⋯, 19, 25]
group2 = [30, 32, ⋯, 39, 35]
group3 = [42, 44, ⋯, 43, 35]

anova, pvalue = f_oneway(group1, group2, group3)

print(“The ANOVA value is “, anova)
print(“The p-value is “, pvalue)
```

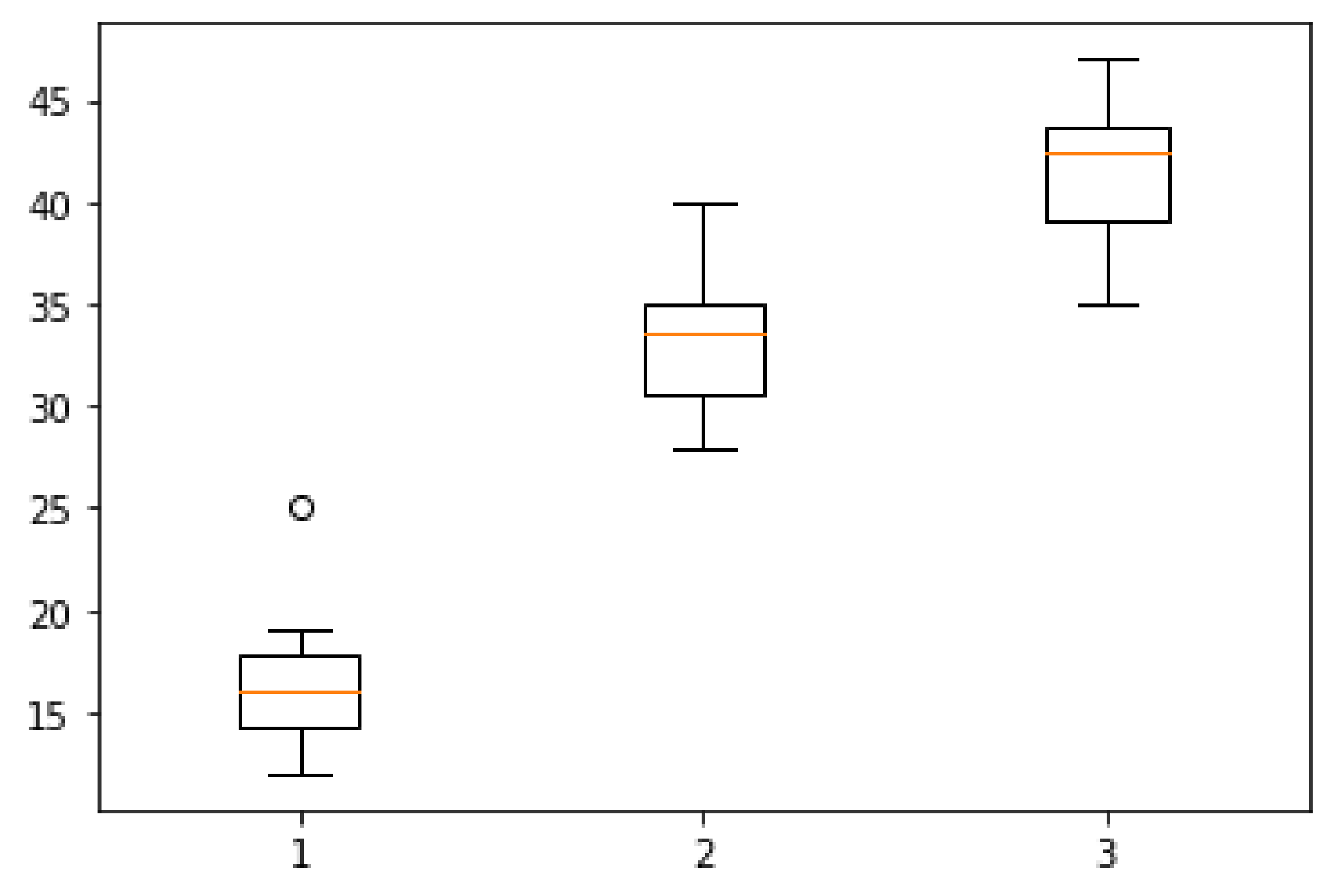
# Post-Hoc test

- It is normally use post-hoc test when reject null hypothesis at ANOVA test
- It is the test to find difference on each population
- The most popular post-hoc test is Tukey’s test

# Post-Hoc test in Python

```
import pandas as pd
import numpy as np
from scipy.stats import f_oneway
from statsnmodels.stats.multicomp import pairwise_tukeyhsd
df = pd.DataFrame({‘score’: [15, 12, …, 19, 25,
                             30, 32, …, 39, 35,
                             42, 44, …, 43, 35],
                   ‘group’: np.repeat([‘group1’, ‘group2’, ‘group3’], repeats=10)})
tukey = pairwise_tukeyhsd(endog = df[‘score’], groups = df[‘group’], alpha = 0.05
print(tukey)
```

# Post-Hoc Test



# One-Way ANOVA practice

Social media: The researcher would like to find the difference of average spending time (minutes) on each social network platform (Facebook, YouTube, and Tiktok). They randomly selected 10 people and collect their spending time. The data is illustrated as the following table:

| Facebook | 34 | 24 | 31 | 29 | 30 | 28 | 32 | 26 | 37 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|
| YouTube | 84 | 91 | 78 | 79 | 82 | 88 | 85 | 81 | 90 | 85 |
| Tiktok | 52 | 54 | 43 | 49 | 48 | 55 | 49 | 57 | 53 | 55 |

a) Test the equality of variance at 5%

b) Test the difference of average spending times on each social network at 5%

c) Test 5% significance level at Tukey HSD and draw box-plot

# Two-Way ANOVA

- It's mostly similar to one-way ANOVA but it used to estimate the mean of a quantitative variable changes according to the levels of two categorical variables.
- However, it's require to test the null and alternative hypothesis in both categorical variables and interactive variable. Hence, they can state as the following:

$H_0$: There are no differences between the first / second categorical variable

$H_a$: At least one population is difference at first / second categorical variable

$H_0$: There is no interaction between first and second categorical variable

$H_a$: There is interaction between first and second categorical variable

# Two-Way ANOVA example

Football game: The researcher would like to test the difference goal in football match. They decide to collect the data from the first 3 week of the league (first, second and third) from each team (home and away). The data is provided as the following table:

| Week | Home | Away |
|---|---|---|
| First | 23, 24, 22, 28, 30 | 14, 15, 18, 13, 20 |
| Second | 31, 33, 28, 30, 25 | 21, 22, 25, 21, 19 |
| Third | 35, 34, 32, 30, 28 | 26, 24, 21, 25, 20 |

Test 5% significance level at both factors and interaction

# Two-Way ANOVA in Python

```
import numpy as np
import pandas as pd
df = pd.DataFrame({‘week’: np.repeat([‘first’, ‘second’, ‘third’], 10),
                   ‘match’: np.tile(np.repeat([‘home’, ‘away’], 5), 3),
                   ‘score’: [23, 24, 22, 28, 30, 14, 15, 18, 13, 20,
                             31, 33, 28, 30, 25, 21, 22, 25, 21, 19,
                             35, 34, 32, 30, 28, 26, 24, 21, 25, 20]})
import statsmodels.api as sm
from statsmodels.formula.api import ols
model = ols(‘score ~ C(week) + C(match) + C(week):C(match)’, data = df).fit()
print(sm.stats.anova_lm(model, typ = 2)
```

You can
type “~”
by Alt+126

# Two-Way ANOVA practice

Computer case: Programmer would like to compare the execution speed (in seconds) for a standard software packages by using three different compilers. Programmer tests each compiler using three different computer models, and the data is shown on the provided table:

| Computer | Compiler | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Model 1 | 9.9 | 8.0 | 7.1 |
| Model 2 | 12.5 | 10.6 | 9.1 |
| Model 3 | 10.8 | 9.0 | 7.8 |

Test at 5% significance level to find the difference on each computer and compiler