Analysis of Variance for Completely Randomized Design

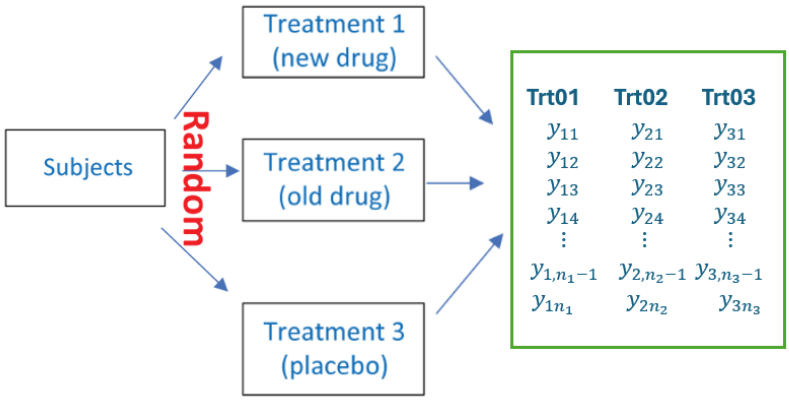
Cheng Peng

STA200 Statistics II

Contents

1 Introduction

Experimental design is a fundamental aspect of statistical analysis, allowing researchers to systematically investigate the effects of different treatments on a response variable. In CRD, experimental units are randomly assigned to different treatment groups.



In this note, we will explore how Classical Analysis of Variance (ANOVA) and Regression Analysis can be used to analyze data from a CRD. We will also illustrate these methods using some numerical examples. For

Working Data in Plant Physiology

Our example data are from an experiment in plant physiology, published by Sokal and Rohlf (1995). The lengths of pea sections (the dependent, or response, variable) grown in a tissue culture were recorded. The purpose of the experiment was to test the effects of various sugar media (the independent, or explanatory, variable) on mean pea section length. A **balanced CRD** was used with 10 replicates per treatment level.

The data are given in the table below:

Observed length of pea sections (ocular units) for the different sugar treatments.

Treatment				
Control	Glucose *2% glucose added*	Fructose *2% fructose added*	GlucFruc *1% glucose+1% fructose added*	Sucrose *2% sucrose*
75	57	58	58	62
67	58	61	59	66
70	60	56	58	65
75	59	58	61	63
65	62	57	57	64
71	60	56	56	62
67	60	61	58	65
67	57	60	57	65
76	59	57	57	62
68	61	58	59	67

The wide-format table is easy to use when performing manual calculations. Most of the software programs use long-format data tables. We will reshape this wide table when using the R program for various analyses.

2 Objective and Logic of Analyzing CRD Data

Before moving to data analysis, we first discuss the objective, logic, and methods for analyzing CRD data. The

2.1 The objective and Logic

The core goals when analyzing Completely Randomized Design (CRD) data are threefold:

- **Treatment Effect Detection** - Determine if different treatments produce statistically significant differences in the response variable. For example, do fertilizers A, B, and C result in different crop

yields?

- **Effect Size Quantification** - Measure the magnitude of treatment differences. For example, Fertilizer C increases yield by 13 bushels/acre compared to Fertilizer A
- **Ranking/Comparison of Treatments** Identify which treatments perform best/worst. This involves multiple comparisons among means across treatments. For example, we can order fertilizers by yield: A > B > A through multiple comparisons.

Statistically, **treatment Effect Detection** in CRD with k treatments can be achieved by testing a hypothesis.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{vs} \quad H_a : \text{ at least one mean differs}$$

If H_0 is **rejected**, we quantify the size of treatment effects and rank the treatment effects across the treatments in the CRD. There are different approaches to testing the above hypothesis

2.2 Measuring Discrepancy between H_0 and H_a

Let's use the following balanced CRD data table and related notations.

	Treatments				
	Trt.1	**Trt.2**	**Trt.3**	⋯	**Trt.t**
1	y_{11}	y_{21}	y_{31}	⋯	y_{t1}
2	y_{12}	y_{22}	y_{32}	⋯	y_{t2}
3	y_{13}	y_{23}	y_{33}	⋯	y_{t3}
4	y_{14}	y_{24}	y_{34}	⋯	y_{t4}
5	y_{15}	y_{25}	y_{35}	⋯	y_{t5}
6	y_{16}	y_{26}	y_{36}	⋯	y_{t6}
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	y_{1n_1}	y_{2n_2}	y_{3n_3}	⋯	y_{tn_t}
Column mean	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$	$\bar{y}_{3\cdot}$	⋯	$\bar{y}_{t\cdot}$

Notations:

Sample size: $n_T = n_1 + n_2 + \cdots + n_t$
Column Mean: $\bar{y}_{1\cdot} = (y_{11} + y_{12} + y_{13} + \cdots + y_{1n_1})/\, n_1$
Grand Total: $\bar{y}_{\cdot\cdot} = \frac{1}{n_T}\sum_{j=1}^{t}\sum_{i=1}^{n_j} y_{ij}$

Three Different Errors

Between Treatment: $\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$
Within i-th Treatment: $y_{ij} - \bar{y}_{i\cdot}$
Residual: $y_{ij} - \bar{y}_{\cdot\cdot}$

The three different errors defined above have the following relationship.

$$(y_{ij} - \bar{y}_{\cdot\cdot}) = (y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$$

The above expression simply means that we can **decompose** the overall residual $(y_{ij} - \bar{y}_{\cdot\cdot})$ into within treatment error $((y_{ij} - \bar{y}_{i\cdot})$, also called within residual error) and between treatment error $(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$. Since an error could be positive or negative. Similar to the definition of variance, we square all these errors to obtain squared errors.

After some tedous algebra (i.e., not straightforward at all!), we have the following equation

$$\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

The last term in the above equation has no subscript j; we simplify the last term to get the following equation.

$$\underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2}_{SS_T}=\underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2}_{SS_W}+\underbrace{\sum_{i=1}^{t}n_i(\bar{y}_{i\cdot}-\bar{y}_{..})^2}_{SS_B}$$

where SS_T = **Total Sum of Squared Residuals**, SS_W = **Sum of Squared Errors within treatments** and SS_B = **Sum of Squared Errors between treatments**.

Recall the definition of sample variance

$$s^2=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots_(\,x_n-\bar{x})^2}{n-1}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1},$$

Which is **average squared deviation** (also called **mean squared deviation MSD or mean squared error, MSE**)! The denominator $n-1$ in the above definition reflects the **degrees of freedom** - meaning that number of independent observations (n) minus **one** constratint ($\bar{x}=\sum_{i=1}^{n}x_i/n$, each equation based on the n independent observatiob is considered as a contraint!).

Let's look at SS_B: There are t independent means (one for each treatment) $\{\bar{y}_{1\cdot},\bar{y}_{2\cdot},\bar{y}_{3\cdot},\cdots,\bar{y}_{t\cdot}\}$, the grand total $\bar{y}_{..}$ is consider a constraint. Therefore, there are $ t-1$ degrees of freedom associated with SS_B. Therefore, the mean squared error of the between errors is defined by

$$MS_B=\frac{SS_B}{t-1}$$

Similarly, in SS_W, there n_T independent observations y_{ij} and t constraints $\{\bar{y}_{1\cdot},\bar{y}_{2\cdot},\bar{y}_{3\cdot},\cdots,\bar{y}_{t\cdot}\}$ (**Yes! since each of them is the mean of observations within each treatment**). Therefore, the degrees of freedom in SS_W is n_T-t, Therefore, the mean square error of within errors is defined by

$$MS_W=\frac{SS_W}{n_T-t}.$$

Remarks:

- MS_B is the estimated sample variance of the **sample means** of individual treatments.
- MS_W is the weighted average of the variances of each individual treatment.

Here is the logic for how to define a measure to assess the discrepancy between H_0 and H_a:

If there is no real difference between treatments (i.e., H_0 is true), MS_B and MS_W both estimate the same underlying variance. Therefore $MS_B\approx MS_W$!

We can construct the test statistic based on MS_B and MSE_B under H_0. Mathematically, both MS_B-MS_W (≈ 0) and MS_B/MS_W (≈ 1) are valid measures to assess the discepancy between H_0 and H_a. Since both quantities are random (since they are evaluated based on the random sample), we need to choose the one with a known **probability distribution**. The ratio expression MS_B/MS_W has a well-known distribution - **F distribution**!

2.3 Basics of F Distribution

We have discussed several distributions to characterize test statistics for the z-test, t-test, and chi-square test. One of the essential tools in hypothesis testing involving the comparison of variances (the case we introduced above) is the F-distribution. This subsection outlines the basics of the F-distribution and demonstrates how to use R to compute critical values and p-values associated with F-tests.

The F-distribution is a continuous probability distribution used for the comparison of two sample variances (such as SS_B and SS_W) and regression analysis (will discuss this later). Using the above quantities, we have

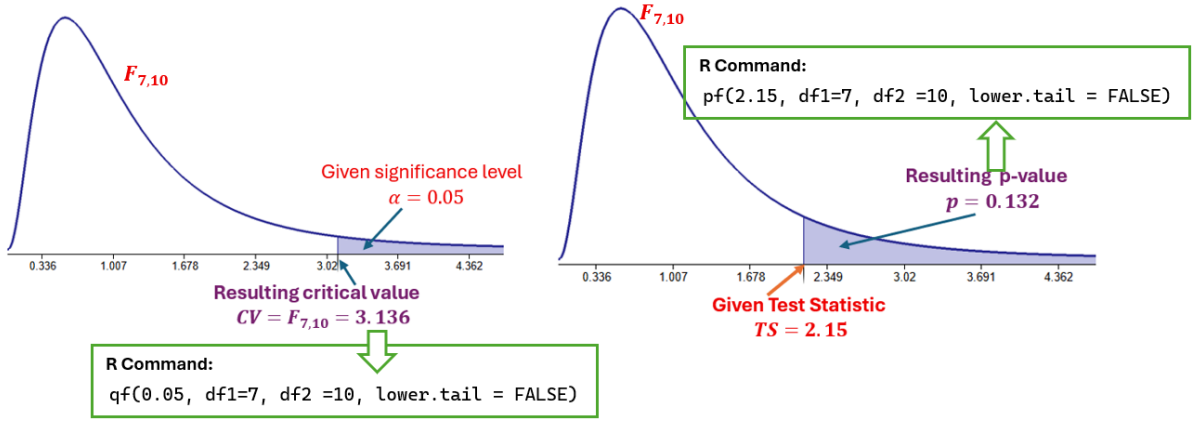
$$F = \frac{SS_W/(n_T - t)}{SS_B/(t-1)} \rightarrow F_{n_T - t,\ t-1}$$

$n_T - t$ is the degrees of freedom of the **numerator** and $t - 1$ is the degrees of the **denominator**. We can also flip the ratio to get another valid test statistic

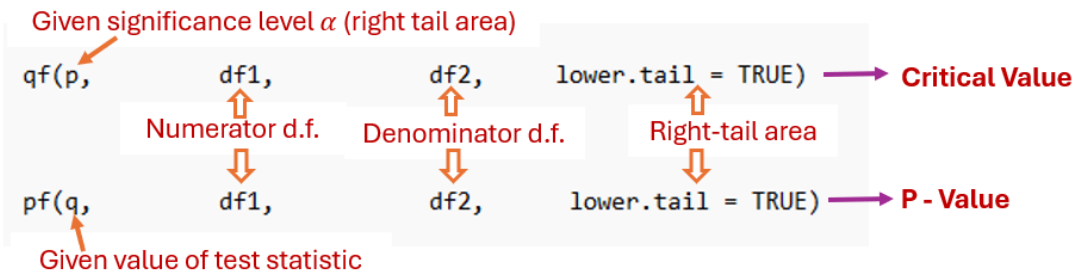
$$F' = \frac{SS_B/(t-1)}{SS_W/(n_T - t)} \rightarrow F_{\ t-1, n_T - t}$$

Caution: the degrees of freedom in the subscript MUST be adjusted so that the first index is the degrees of freedom of the numerator and the second subscript represents the degrees of freedom of the denominator. **This means $F_{7,5}$ and $F_{5,7}$ are two different F distributions!**

F_{df_N, df_D} is skewed to the right. The following figure shows the density curve of an F-distribution with degrees of freedom df1 = 7 (numerator) and df2 = 10 (denominator), together with the R commands to find the critical and p-values.



The syntax of the two R functions used to find the p-value and critical value is annotated in the following figure.



Most introductory statistics textbooks include an F table based on a few commonly used significance levels α. For example, the critical value of $F_{7,10}$ at $\alpha = 0.05$ labeled in the left panel of the density curve of $F_{7,10}$ can also be found from the following F table

F-table of Critical Values of α = 0.05 for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07

The F tables are constructed based on a few significance levels. It is not to use use F table to find p-values.

Next, we open a standalone section to formally introduce the one-way ANOVA for CRD data.

3 One-Way ANOVA and Implementation

We have discussed the logic and valid measures for assessing the discrepancy between H_0 and H_a. Next, we summarize the above results in a classical one-way ANOVA table and test the following null hypothesis based on the CRD.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t \quad \text{vs} \quad H_a : \text{at least one mean differs}$$

The test statistic and corresponding p-value are summarized in the following ANOVA table.

Source	**DF**	**SS**	**MS**	**F-Value**	**P-Value**
Between	$t-1$	SS_B	$MS_B = SS_B/(t-1)$	MS_W/MS_B	
Within	$n_T - t$	SS_W	$MS_W = SS_W/(n_T - t)$		
Total	$n_T -$ ‘	SS_T			

Decision Rule: If the p-value in the above table is less than the significance level α, we reject the null hypothesis; otherwise, we conclude the null hypothesis.

Next, we use the **plant physiology data** to demonstrate to to construct the above ANOVA table and use R as a calculator to find SS and MS and related quantities in the ANOVA table. The following code loads the data into R in the same format (i.e., wide table format)

```
# Create the data vectors
Control <- c(75, 67, 70, 75, 65, 71, 67, 67, 76, 68)      # Control
Glucose <- c(57, 58, 59, 59, 62, 60, 60, 57, 59, 61)      # Glucose
```

```
Fructose <- c(58, 61, 56, 58, 57, 56, 58, 60, 57, 58)     # Fructose
GlucFruc <- c(58, 59, 58, 61, 57, 56, 58, 57, 57, 59)     # GlucFruc
Sucrose <- c(62, 66, 65, 63, 64, 62, 65, 65, 62, 67)     # Sucrose
## using cbind() to display the data
cbind(Control, Glucose, Fructose, GlucFruc, Sucrose)
```

```
|       Control Glucose Fructose GlucFruc Sucrose
|  [1,]      75      57       58       58      62
|  [2,]      67      58       61       59      66
|  [3,]      70      59       56       58      65
|  [4,]      75      59       58       61      63
|  [5,]      65      62       57       57      64
|  [6,]      71      60       56       56      62
|  [7,]      67      60       58       58      65
|  [8,]      67      57       60       57      65
|  [9,]      76      59       57       57      62
| [10,]      68      61       58       59      67
```

```
#colnames(wide.table) <- c("Control", "Glucose", "Fructose", "GlucFruc", "Sucrose")
```

The next code chunk calculates different quantities in the ANOVA table. I will make comments in the code to show detailed steps. The primary R commands are `sum()`, `mean()`, and related R functions of the F-distribution.

```
## degrees of freedom
t = 5
n.T = 10*5
dfN = t - 1
dfD = n.T - t
ybar.. = mean(c(Control, Glucose, Fructose, GlucFruc, Sucrose))
## Sum of squares
SS.B = 10*sum(c((mean(Control)-ybar..)^2, (mean(Glucose)-ybar..)^2, (mean(Fructose)-ybar..)^2, (mean(Glu
SS.W = sum(c(sum((Control-mean(Control))^2), sum((Glucose-mean(Glucose))^2), sum((Fructose-mean(Fructos
SS.T = sum((c(Control, Glucose, Fructose, GlucFruc, Sucrose) - ybar..)^2)
## MS
MS.B = SS.B/dfN
MS.W = SS.W/dfD
## F value
F.value = MS.B/MS.W
p.val = pf(F.value, dfN, dfD, lower.tail = FALSE)
anova.out <- data.frame(
  DF = c(dfN, dfD),
  SS = c(SS.B, SS.W),
  MS = c(MS.B, MS.W),
  F.value = c(F.value, NA),
  p.value = c(p.val, NA)
)
rownames(anova.out)=c("Between", "Within")
print(anova.out, na.print = "")
```

```
|         DF       SS         MS  F.value      p.value
| Between  4 1077.944 269.486000 51.31981 3.324636e-16
| Within  45  236.300   5.251111       NA           NA
```

Conclusion: The p-value is approximately 0. We reject the null hypothesis that all means are equal. In other words, the mean pea section length is affected by various sugar media.

The second part of the following YouTube video (after 15 Minutes, https://www.youtube.com/watch?v=KJ5G2KjcXcA) discussed in the next note on One-way ANOVA.

4 ANOVA with R Function `aov()`

We still use **plant physiology data** to illustrate one-way ANOVA analysis using `aov()`. Since `aov()` requires a long table structure. That is, all measurements of the response must be stored in a column and a standalone column to label the treatment of each value in the measurement column. The following code creates a long table.

```
# Create the data vectors
lengths <- c(
  75, 67, 70, 75, 65, 71, 67, 67, 76, 68,     # Control
  57, 58, 59, 59, 62, 60, 60, 57, 59, 61,     # Glucose
  58, 61, 56, 58, 57, 56, 58, 60, 57, 58,     # Fructose
  58, 59, 58, 61, 57, 56, 58, 57, 57, 59,     # GlucFruc
  62, 66, 65, 63, 64, 62, 65, 65, 62, 67      # Sucrose
)

# Create the treatment factor
treatment <- rep(c("Control", "Glucose", "Fructose", "GlucFruc", "Sucrose"), each = 10)

# Combine into a data frame
pea.long <- data.frame(
  Length = lengths,
  Treatment = factor(treatment, levels = c("Control", "Glucose", "Fructose", "GlucFruc", "Sucrose"))
)

# View the first 20 rows to make sure the data structure is correct
head(pea.long, n=20)
```

```
|    Length Treatment
| 1      75   Control
| 2      67   Control
| 3      70   Control
| 4      75   Control
| 5      65   Control
| 6      71   Control
| 7      67   Control
| 8      67   Control
| 9      76   Control
| 10     68   Control
| 11     57   Glucose
| 12     58   Glucose
| 13     59   Glucose
| 14     59   Glucose
| 15     62   Glucose
```

```
| 16      60   Glucose
| 17      60   Glucose
| 18      57   Glucose
| 19      59   Glucose
| 20      61   Glucose
```

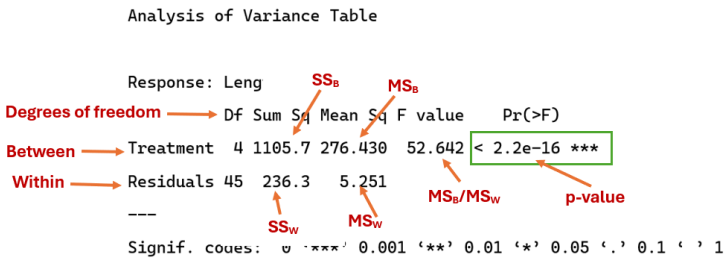
We next call `aov()` to perform ANOVA analysis.

```
## create an anova model object
aov.model <- aov(Length ~ Treatment, data = pea.long)
## create the classical ANOVA table
anova(aov.model)
```

```
| Analysis of Variance Table
|
| Response: Length
|           Df Sum Sq Mean Sq F value    Pr(>F)
| Treatment  4 1105.7 276.430  52.642 < 2.2e-16 ***
| Residuals 45  236.3   5.251
| ---
| Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## or alternatively, you can simply use summary(aov.model) to produce the same results!
```

The annotated output is given by



Conclusion: Since the p-value is approximately 0, the null hypothesis is rejected. That is, the various sugar media (the independent, or explanatory, variable) significantly affect the mean pea section length.

The following YouTube video (https://www.youtube.com/watch?v=IkR3Rzrgiv4) gives another example of using the R function `aov()`. The video also uses some graphical functions to create some visualizations. We did not discuss visualizations in this class, but we will cover more visualizations in subsequent statistics courses.

5 Multiple Comparisons

When analyzing variance (ANOVA), a significant F-test (i.e., p-value is less than the given significance level α) indicates that at least one group mean differs from the others. However, ANOVA does not specify which pairs of groups are significantly different. **Post-hoc tests**, also known as **multiple comparison tests**,

are used to identify these specific differences while controlling for the increased risk of Type I errors (false positives) that arise from performing multiple comparisons.

There are different **Post-hoc tests**. We only introduce two commonly used post-hoc methods using R built-in functions:

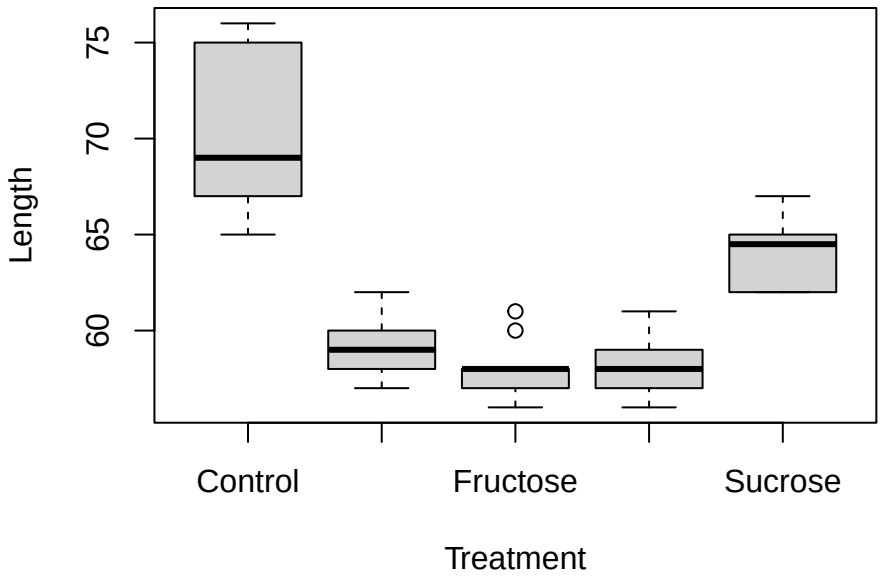
- **Tukey's Honest Significant Difference (HSD)** -Controls **family-wise error rate** (FWER) for all pairwise comparisons.
- **Bonferroni Correction** - Adjusts p-values by multiplying them by the number of comparisons

6 Visual Comparison

We have learned different graphical summaries of data. Box-plot is a geometric representation of the 5-number-summary of a numerical variable. Since treatment variable has multiple categories, we can create a back-to-back box-plot to visualize the different in the distributions.

The R function `boxplot()` can draw multiple boxplots of a numerical variable according the categories of the group variable. Since package **{graphics}** comes with the base R, installation and loading the package is unnecessary.

```
boxplot(Length ~ Treatment, data = pea.long)
```



The above box-plots clearly show that the pea lengths using sugar media Fructose, Glucose, and GlucFruc are similar to each other.

6.1 Tukey's Honest Significant Difference (HSD)

Tukey's Honest Significant Difference (HSD) is a post-hoc test used after a statistically significant ANOVA to determine which specific group means differ from each other. Unlike multiple t-tests, which inflate the Type I error rate, Tukey's HSD adjusts for multiple comparisons, maintaining the family-wise error rate (FWER) at the desired level (e.g., 0.05).

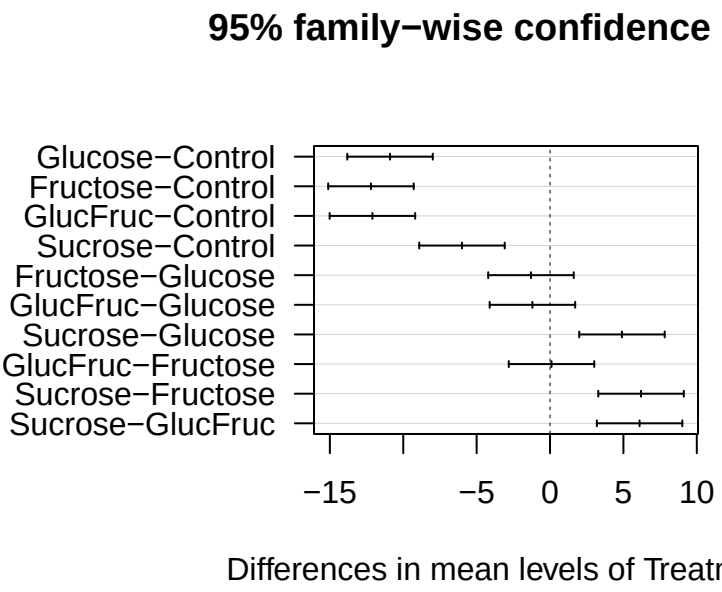
Next, we perform a multiple comparison of pea length across different sugar media using the **plant physiology data**. For convenience, we use the long-format data frame defined earlier and an anova object in the following R code.

```
## refit the ANOVA model
aov.model <- aov(Length ~ Treatment, data = pea.long)
## multiple comparison
tukey.comp <- TukeyHSD(aov.model)
print(tukey.comp)
```

```
|   Tukey multiple comparisons of means
|     95% family-wise confidence level
|
| Fit: aov(formula = Length ~ Treatment, data = pea.long)
|
| $Treatment
|                     diff        lwr       upr     p adj
| Glucose-Control   -10.9 -13.811928 -7.988072 0.0000000
| Fructose-Control  -12.2 -15.111928 -9.288072 0.0000000
| GlucFruc-Control  -12.1 -15.011928 -9.188072 0.0000000
| Sucrose-Control    -6.0  -8.911928 -3.088072 0.0000050
| Fructose-Glucose   -1.3  -4.211928  1.611928 0.7114379
| GlucFruc-Glucose   -1.2  -4.111928  1.711928 0.7676225
| Sucrose-Glucose     4.9   1.988072  7.811928 0.0001778
| GlucFruc-Fructose   0.1  -2.811928  3.011928 0.9999786
| Sucrose-Fructose    6.2   3.288072  9.111928 0.0000026
| Sucrose-GlucFruc    6.1   3.188072  9.011928 0.0000036
```

The results indicate that there is no significant difference in pairs Fructose-Glucose, GlucFruc-Glucose, and GlucFruc-Fructose.

```
par(mai=c(1.5,2,1,1)) # Makes room on the plot for the group names
plot(tukey.comp,      # name of the TukeyHSD() object
     cex.lab = 0.6,   # adjust the font size of the labels of the vertical axis
     las = 1)         # orientation of the labels
```



6.2 Bonferroni Comparison

The Bonferroni correction is a conservative method for adjusting significance levels when performing multiple comparisons to reduce the risk of Type I errors (false positives). It is one of the simplest and most widely used approaches for controlling the family-wise error rate.

The R function `pairwise.t.test()` performs pairwise comparison. It requires inputting individual variables (response and treatment group). The following is code for the Bonferroni procedure.

```
Bonfeeroni.result <- pairwise.t.test(
  x = pea.long$Length,
  g = pea.long$Treatment,
  p.adjust.method = "bonferroni"
)
print(Bonfeeroni.result)
```

```
|
|   Pairwise comparisons using t tests with pooled SD
|
| data:  pea.long$Length and pea.long$Treatment
|
|          Control Glucose Fructose GlucFruc
| Glucose  7.3e-13 -       -        -
| Fructose 1.7e-14 1.00000 -        -
| GlucFruc 2.2e-14 1.00000 1.00000  -
| Sucrose  5.1e-06 0.00019 2.6e-06  3.7e-06
|
| P value adjustment method: bonferroni
```

Using the p-value adjusted Bonferroni procedure, three pairs of treatment groups are significantly different: Fructose-Glucose, GlucFruc-Glucose, and GlucFruc-Fructose. This result is consistent with Tukeu's HSD

method.

Before conclude this section, we recommend a YouTube video (https://www.youtube.com/watch?v=hPUvqtmCu7Q) with an example one multiple comparison among group means using R,

7 Concluding Remarks

One-Way ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more independent groups to determine if at least one group mean is significantly different from the others. It extends the t-test (which compares only two groups) to multiple groups.

Hypotheses

- **Null Hypothesis (H_0)**: All group means are equal($\mu_1 = \mu_2 = \cdots = \mu_k$).
- **Alternative Hypothesis (H_a)**: At least one group mean is different.

The fundamental assumptions of One-Way ANOVA include:

- **Independence**: Observations must be independent (e.g., random sampling).
- **Normality**: Data in each group should be approximately normally distributed (checked via Q-Q plots or Shapiro-Wilk test).
- **Homogeneity of Variance** (Homoscedasticity): Groups should have equal variances (tested using Levene's test or Bartlett's test).

If any of theses assumptions are violated, the classical ANOVA results will be invalid. There some alternatives of the classic ANOVA that will be covered in the subsequent related courses. The key information in the ANOVA analysis is summarized in the following one-way ANOVA table.

Source	df (Degrees of Freedom)	SS (Sum of Squares)	MS (Mean Square)	F-Statistic
Between Groups	$k-1$	SSB	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups (Error)	$N-k$	SSW	$MSW = \frac{SSW}{N-k}$	-
Total	$N-1$	SST	-	-

If ANOVA rejects the null hypotheis, post-hoc tests identify which specific groups differ. We introduced two prcedures for this purpose.

- **Tukey's HSD** (controls family-wise error rate).
- **Bonferroni Correction** (conservative adjustment).