Introduction to Statistics B

Maria Anastasiadi

2025-04-25

Statistical Foundations Part2

4. Analysis of Variance (ANOVA)

Introduction

- A limitation with the t-test is that only two means can be compared at one time.

- However, in many experimental set-ups we want to compare more than two means simultaneously.

- Testing each pair of means with a t-test is not recommended as the probability of false positives increases with each test run.

ANOVA Definiftion

Analysis of Variance (ANOVA) is the recommended method for determining whether or not there is a statistically significant difference between the means of three or more independent groups.

 ANOVA answers the question as to whether there is greater variability between groups than within groups.

ANOVA Hypothesis

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

  The means are equal for each group.

- H_1: at least one of the means is different from the others

 ANOVA models facilitate the analysis of many different kinds of experimental data and they are the workhorse of basic statistical analysis.

4.1 One-Way ANOVA

One-way ANOVA predicts how the mean value of a numeric variable (the **response variable**) is affected by the levels of a categorical variable (the **predictor variable**).

These levels may represent:

a) quantitative variations (e.g. the effect of different concentrations of an antibiotic on bacterial growth).

b) qualitative variations (e.g. the effect of apple cultivar on sugar/acid ratio).

ANOVA and Linear Regression

- The definition of ANOVA is similar to the definition of simple linear regression you have already encountered earlier.

- **In fact, ANOVA and regression are both special cases of the general linear model.**

4.1.1 ANOVA Principles

- ANOVA examines the magnitudes of three different sources of variation in the data:

A) **The Total Variation**: the variation among all the units in the study.

B) **Between-Group Variation**: the variation due to the effect of experimental treatments or control groups (**explained variation**).

C) **Within-Group Variation**: the variation due to other sources (**error variation**).

ANOVA in a Nutshell

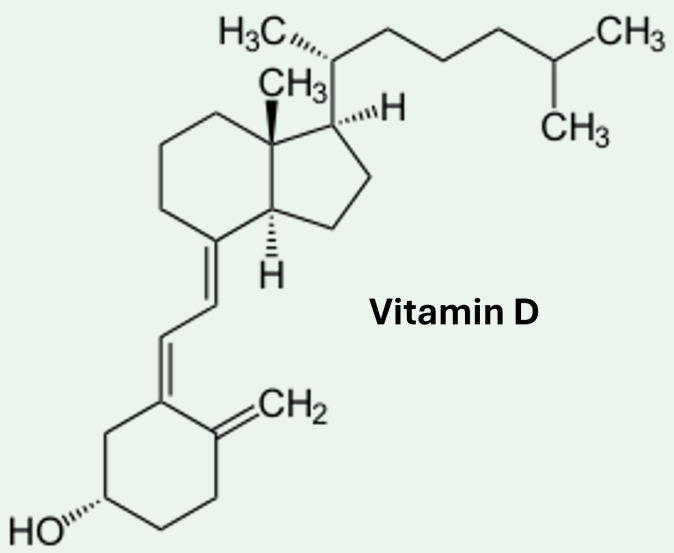
ANOVA is looking at changes in variation. If the amount of variation between treatments is sufficiently large compared to the within-group variation, this suggests that the treatments are probably having an effect.

- But how is each type of variation calculated?

Example

Consider an experiment where we compare the bioaccesibility of Vitamin D depending on the type of flour used in a baked product.

- The fibres used are :wheat pea, apple,



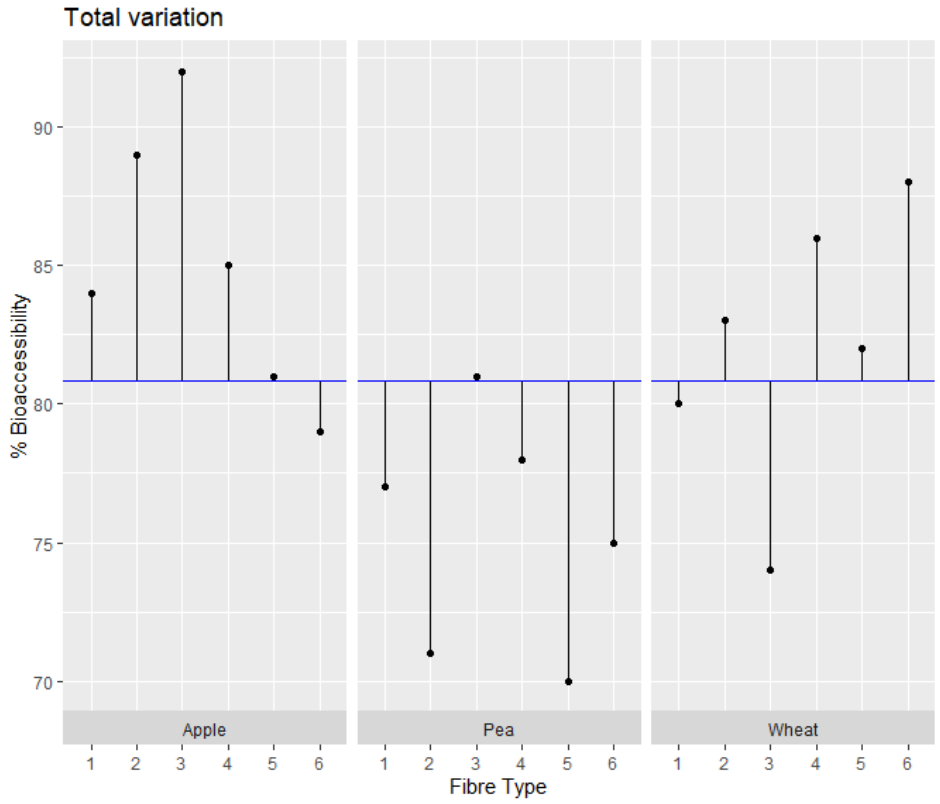
Vitamin D

Import csv file `Fibre` into R

First lines of Fibre dataset

Fibre	Replicate	Bioaccessibility
Wheat	1	80
Wheat	2	83
Wheat	3	74
Wheat	4	86
Wheat	5	82
Wheat	6	88
Pea	1	77
Pea	2	71
Pea	3	81
Pea	4	78

Plot raw data and total mean



- The distance of each sample from the blue line represents the deviation of each measurement from the total mean. The sum of all the deviations is zero.

1. Calculate the Total Variation (SST)

- To find the total variation we need to find the **sum of squares of the deviations (SST)**.

```
## [1] 624.5
```

2. Calculate the Within-Groups Variation (SSE)

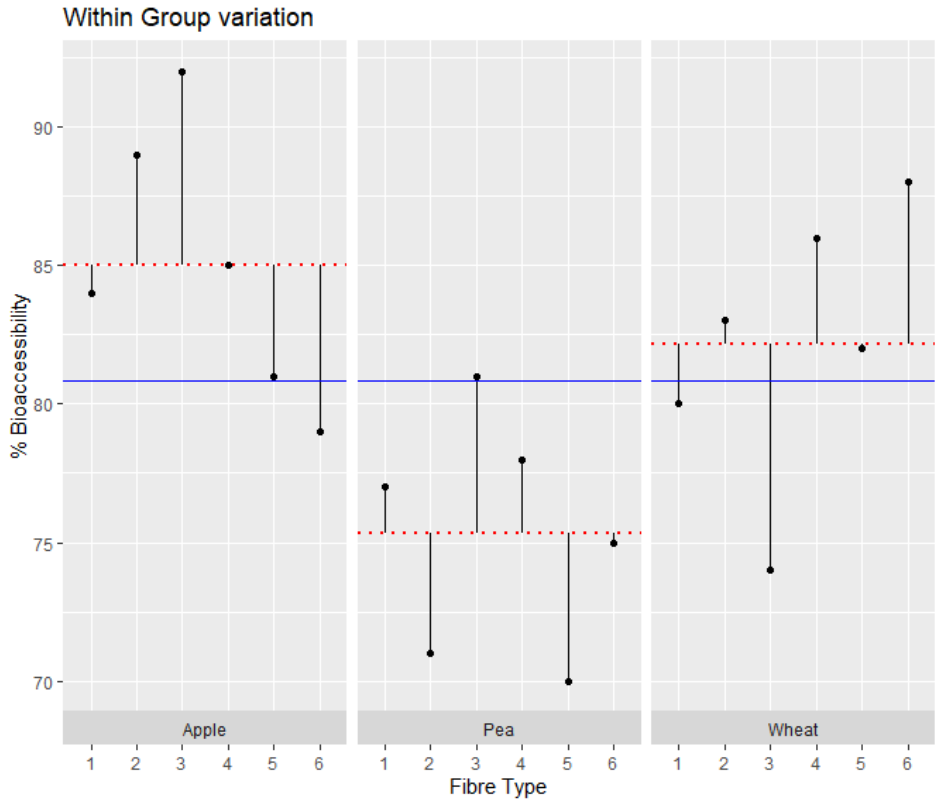
- The next step is to calculate the *error or residual variation*.

- First we need to calculate the means per group.

'group.mean' table

Fibre	**GM**
Apple	85.00000
Pea	75.33333
Wheat	82.16667

Plot raw data, total mean & group means



- The distance of each sample from the red line is the difference of each measurement from the group mean. This is the ‘left over’ variation attributed to differences among individuals.

Within-group variability

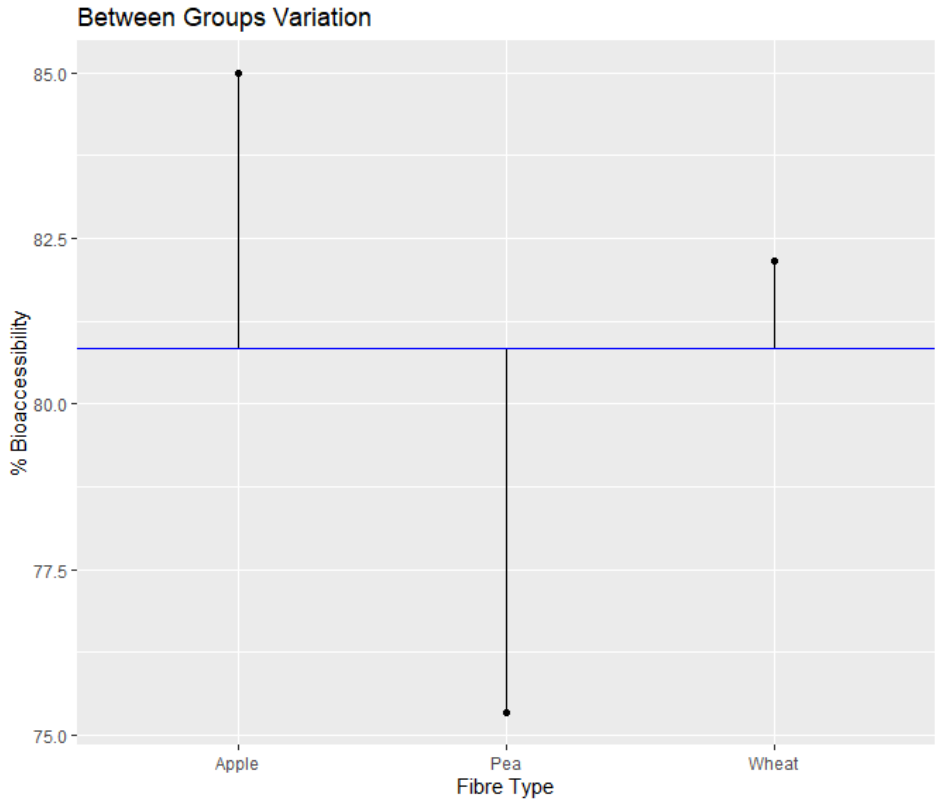
- To calculate the within-group variability we need to take the sum of squares for the deviations from each group mean. This is called the **residual sum of squares**.

Calculate sum of squares (**SSE**)

```
## [1] 328.1667
```

3. Between-Group Variation (SSG)

Plot total mean and group means



- The distance of each dot from the blue line is the difference between each group mean and the Total Mean. This is variation due to differences among treatment groups.

Between Treatments Variability

- As previously we need to calculate the sum of squares for these differences. This is a measure of the variability attributed to differences among treatments.

Calculate sum of squares (**SSG**)

```
## [1] 296.3333
```

Total Sum of Squares

 The SUM of SQUARES we calculated earlier are related by the formula:

$SST = SSG + SSE$

Thus the total variation is composed of two parts, one due to groups and one due to error.

4.1.2 Degrees of Freedom in ANOVA

- An issue with using the sum of squares we calculated previously, is that they are dependent on the sample size and the number of groups.

- To standardise the sum of squares we divide by the degrees of freedom for each type of variation.

Degrees of Freedom Formula

 The Degrees of Freedom in ANOVA are related by the formula:

$$DFT = DFG + DFE$$

DFT = (Number of observations - 1)

DFG = (Number of treatment groups - 1)

DFE = (Number of observations - Number of groups)

4.1.3 Mean Squares

The Sum of Squares for each type of variation in ANOVA is calculated as:

$$MST = \frac{SST}{DFT}$$

$$MSG = \frac{SSG}{DFG}$$

$$MSE = \frac{SSE}{DFE}$$

The Mean Sum of Squares is the standardised form of the Sum of Squares.

4.1.4 The *F* test

- The final question is whether we can reject the Null Hypothesis or not.

- To decide this we use the **ANOVA *F* statistic**.

F Statistic Definition

- The *F* statistic is the ratio of MSG/MSE (the variation due to treatment over the variation due to error).

- If $H_0 =$ TRUE the *F*-statistic is ~ 1 and if the Alternative hypothesis is true, it tends to be large.

The ANOVA *F* test

To test the Null Hypothesis in a One-Way ANOVA we calculate the *F* statistic:

$$F = \frac{MSG}{MSE}$$

F-test P-Value

The **P-value** of the *F* test is the probability that a random variable having the $F(I-1, N-1)$ distribution is $\geq F$, the calculated value of the *F* Statistic.

Example

- Find the *F* statistic for the bioaccessibility problem and decide whether we can reject the Null Hypothesis.

1. Calculate Degrees of Freedom

```
## [1] 17
```

```
## [1] 2
```

```
## [1] 15
```

2. Find MSS

```
## [1] "mst= 36.74"
```

```
## [1] "msg= 148.17"
```

```
## [1] "mse= 21.88"
```

3. Find the *F* statistic

```
## [1] "f=msg/mse= 6.77"
```

The *F* statistic is 6.77.

F Table

- If we look at an F table for critical values for α=0.05, the critical value for *F(2,15)* is 3.68 which is smaller than our F value. So we can reject the Null hypothesis.

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
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96

4.1.5 Coefficient of Determination R^2

- Another statistic we can calculate from an ANOVA table is the **coefficient of determination**

$$R^2 = \frac{SSG}{SST}.$$

```
rsq <- explained.variation/total.variation
rsq
```

```
## [1] 0.4745129
```

- This coefficient tells us that 47.5% of the total variation in Bioaccessibility is explained by the different type of fibre and the other 52.5% is explained by sample-to-sample variation

4.1.6 One Way Anova using R

- Base R can carry out a One-Way ANOVA using simple functions such as the `lm()` and `aov()` functions.

- There are also dedicated statistical libraries such as **afex** which can carry out ANOVA.

How to do a one-way ANOVA in one step:

```
anova1 <- aov(Bioaccessibility~as.factor(fibre$Fibre), data=fibre)
```

Print ANOVA summary

```
summary(anova1)
```

```
##                          Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(fibre$Fibre)  2  296.3  148.17   6.772 0.00802 **
## Residuals               15  328.2   21.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check if the values in the ANOVA summary table match the values we got by doing the calculations manually.

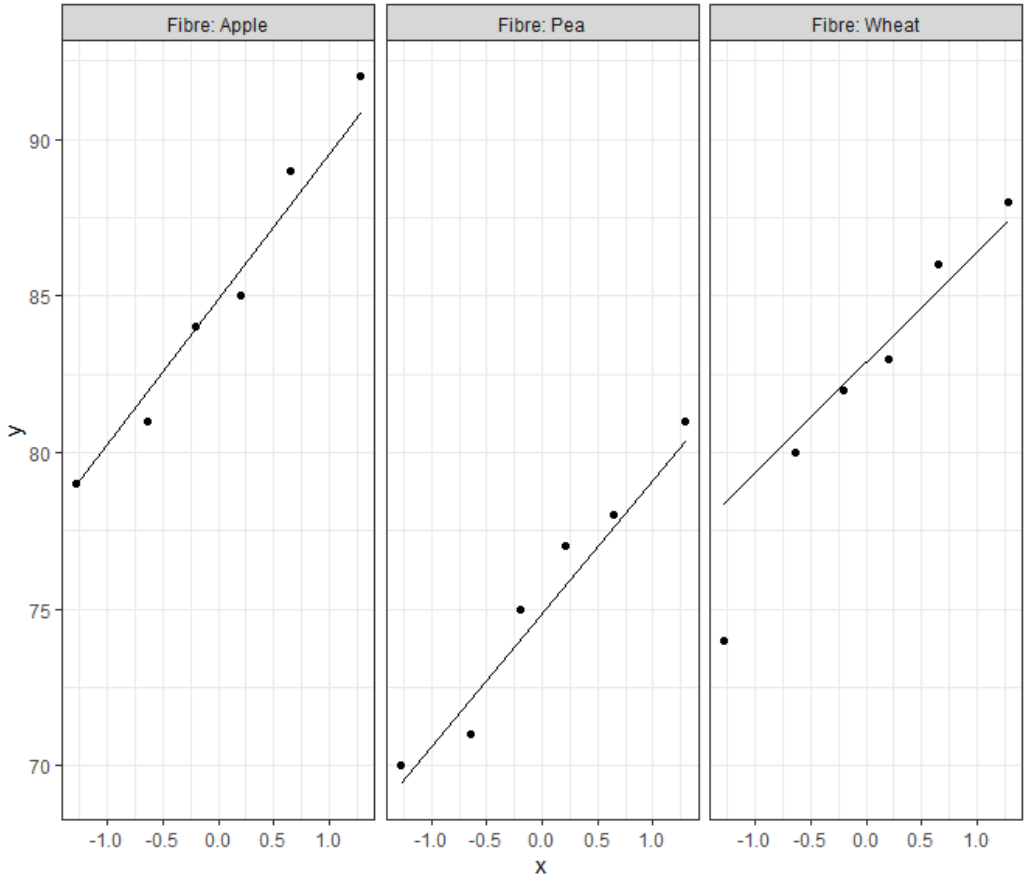
4.1.7 Assumptions for One-Way ANOVA

- The continuous variable has a NORMAL distribution in ALL relevant populations (groups) or at least it doesn't have any gross outliers.

- Not as important if the sample is large (Central Limit Theorem).

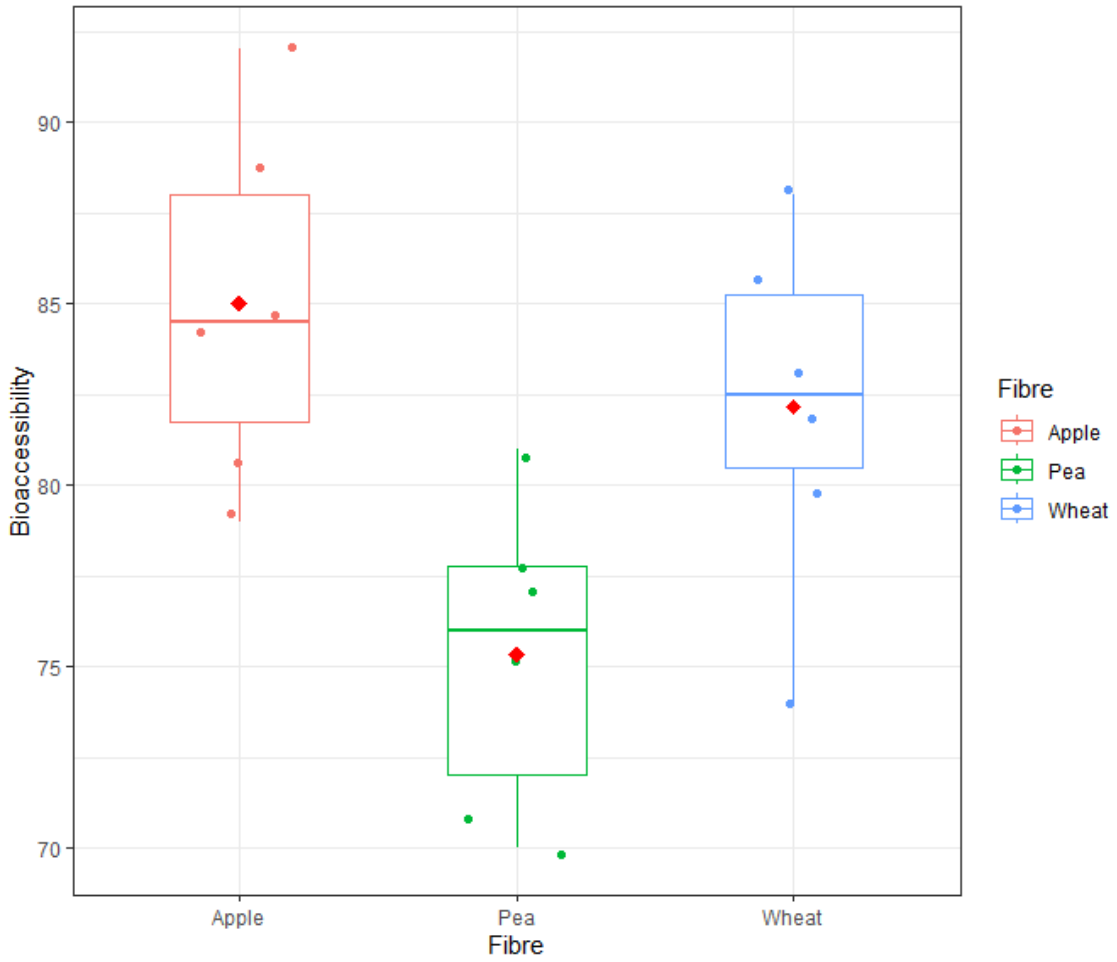
- If the sample is far from normal &/or small, we may need to consider alternative methods (non-parametric).

Create QQ plot for each group



QQ plots sort data in ascending order, and plot them against quantiles from a theoretical normal distribution.

Create a box-plot



4.1.8 Assumptions for the Residuals

The main assumption the residuals need to meet are the following:

Residual Assumptions		
Assumption	**How to Check**	**What to Do if Not Met**
1. Residuals are normally distributed	Check histogram and QQ plots of residuals. Shapiro-Wilk's Test for normality	Consider data transformation (e.g. sqrt or log). If still not normal use a non-parametric test (Kruskall- Wallis)
2. Homogeneity of variance (The variances s^2 should be equal for all groups)	Residuals vs fitted plot. Levene test for equality of variances	If p-value < 0.05, consider data transformation or use the Welch test

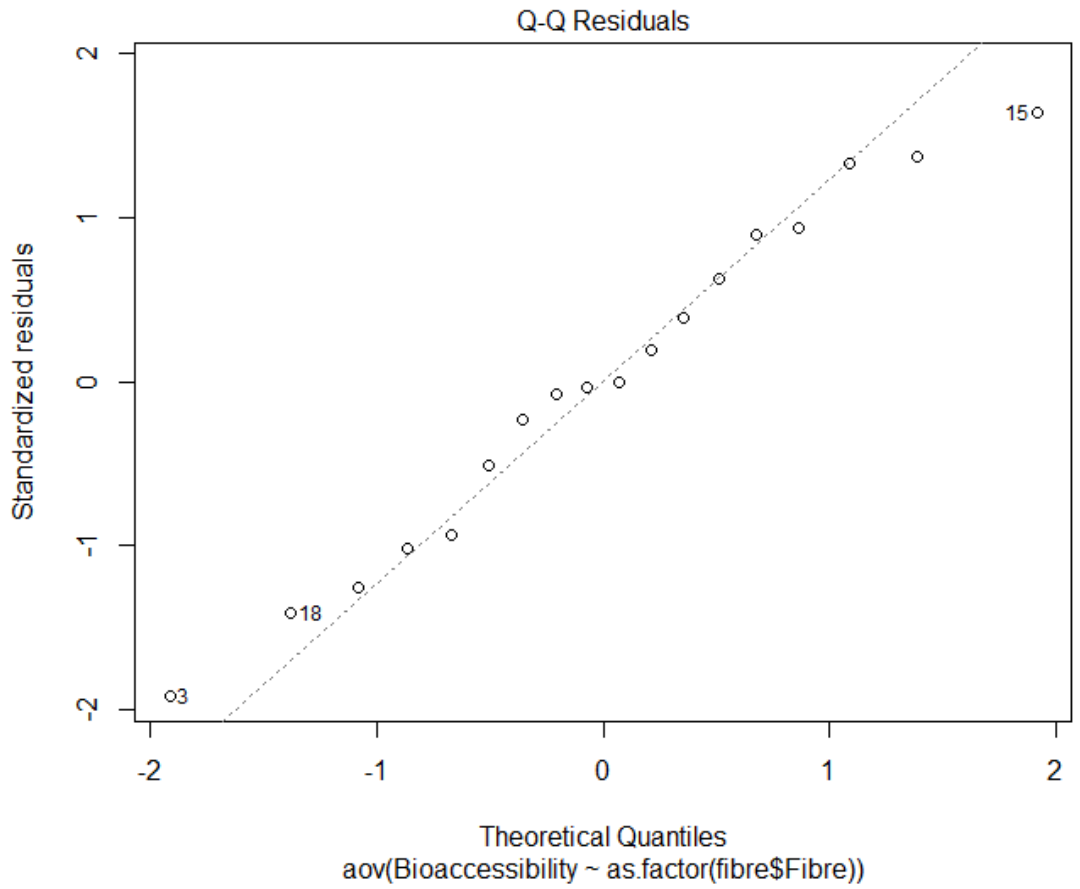
Check Residual Assumptions

Let's check the assumptions for the `fibre` dataset.

1) Residuals QQ plot

```r
## Draw QQ plot
plot(anova1, 2)
```

Visualise residuals QQ plot

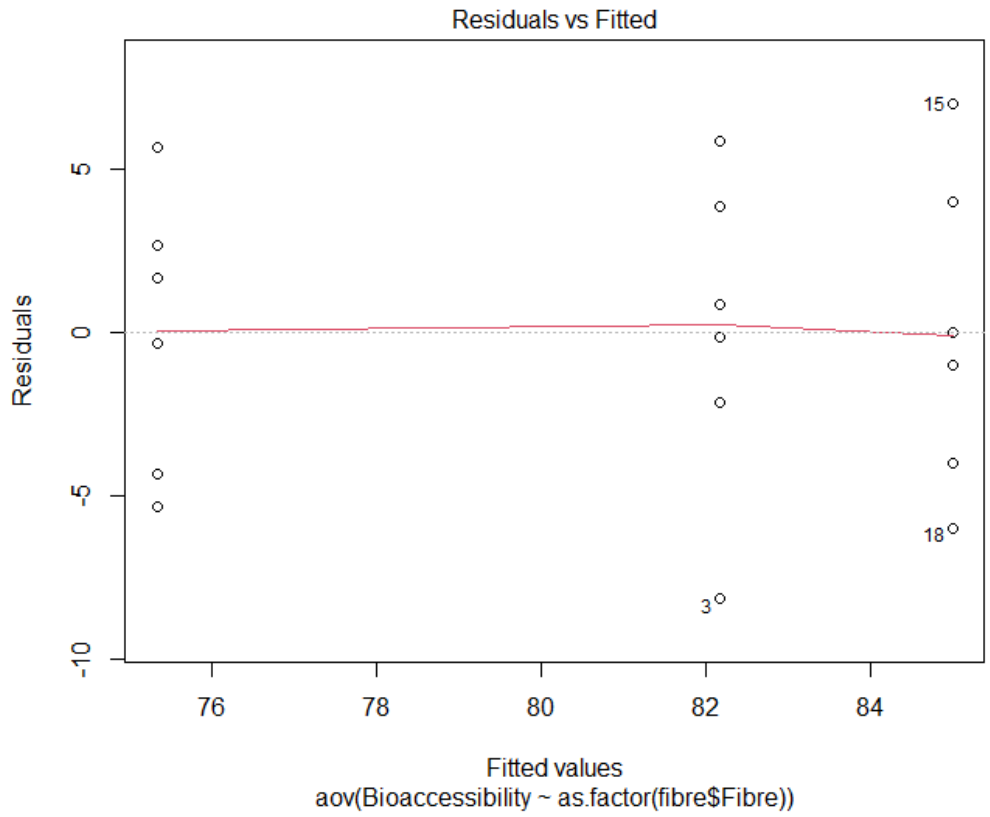


No severe deviations from normality

2) Equality of Variation:

```
## Draw Fitted values vs Residuals plot
plot(anova1, 1)
```

Visualise residual variance plot



Again, we don’t see any deviations from this assumption. To make sure we can apply Levene test.

4.1.9 Multiple Comparisons

In the Bioaccessibility example we rejected the null hypothesis (All Means are Equal) in favour of the Alternative hypothesis (Not All Means are Equal).

▪ This however is not very informative. We want to know which group means are statistically significantly different.

▪ To do this we need to make multiple pairwise comparisons using *t*-tests.

▪ However, when many *t*-tests are applied simultaneously we run the risk of false positives.

4. 1.10 Post-Hoc Tests

- To address the risk of false positives we apply **Post-Hoc** tests (because they can only be applied after we reject H_0)

Most Popular Post-Hoc Tests

- Bonferroni test
- Benjamini-Hochberg test
- Scheffé's test
- Duncan's new multiple range test
- Tukey Honest Significant Differences test (Tukey HSD)

Tukey Test in R

```
TukeyHSD(anova1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Bioaccessibility ~ as.factor(fibre$Fibre), data
= fibre)
##
## $`as.factor(fibre$Fibre)`
##                  diff         lwr       upr     p adj
## Pea-Apple   -9.666667 -16.6810832 -2.652250 0.0072578
## Wheat-Apple -2.833333  -9.8477499  4.181083 0.5586372
## Wheat-Pea    6.833333  -0.1810832 13.847750 0.0567231
```

Which pairs of techniques vary significantly?

4.1.10 Plot the results

- Finally we plot the means for each group in a barplot.
- Before doing that we need to calculate the standard error (se) for each group so we can add it to each bar. Remember we can find the se from the formula: $se = s/\sqrt{n}$

Visualise bar plot for the data