Introduction to Basic R: Statistical Analysis

Microbiome analysis course

Presented by
Trần Bùi Minh Trí
Nguyễn Thanh Trà

June 13 2024

Content

Introduction to statistics

Data in the nutshell

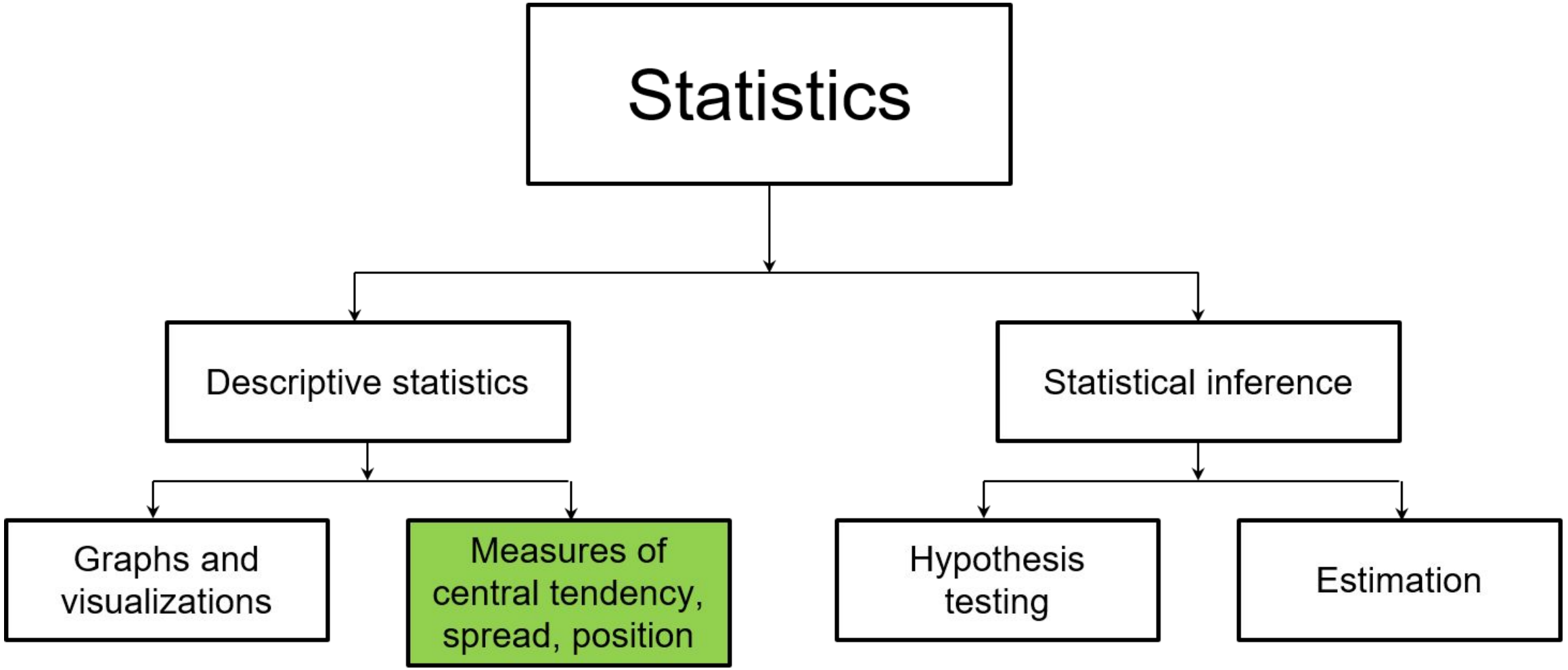
Descriptive statistics

Data distribution

Introduction to hypothesis in microbiome studies

Inferential statistics: Hypothesis testing and linear regression

1. Introduction to statistics



1.Introduction of the dataset

Iris dataset

```
> head(dt.iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

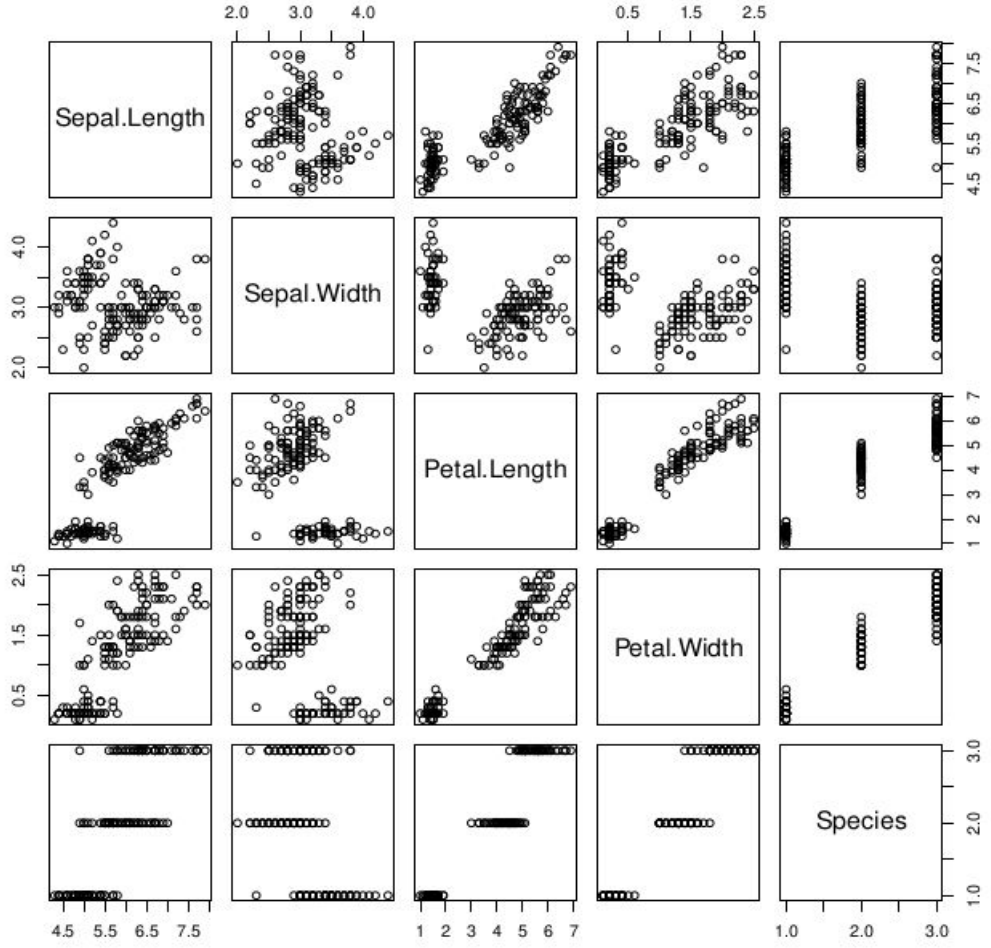
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa

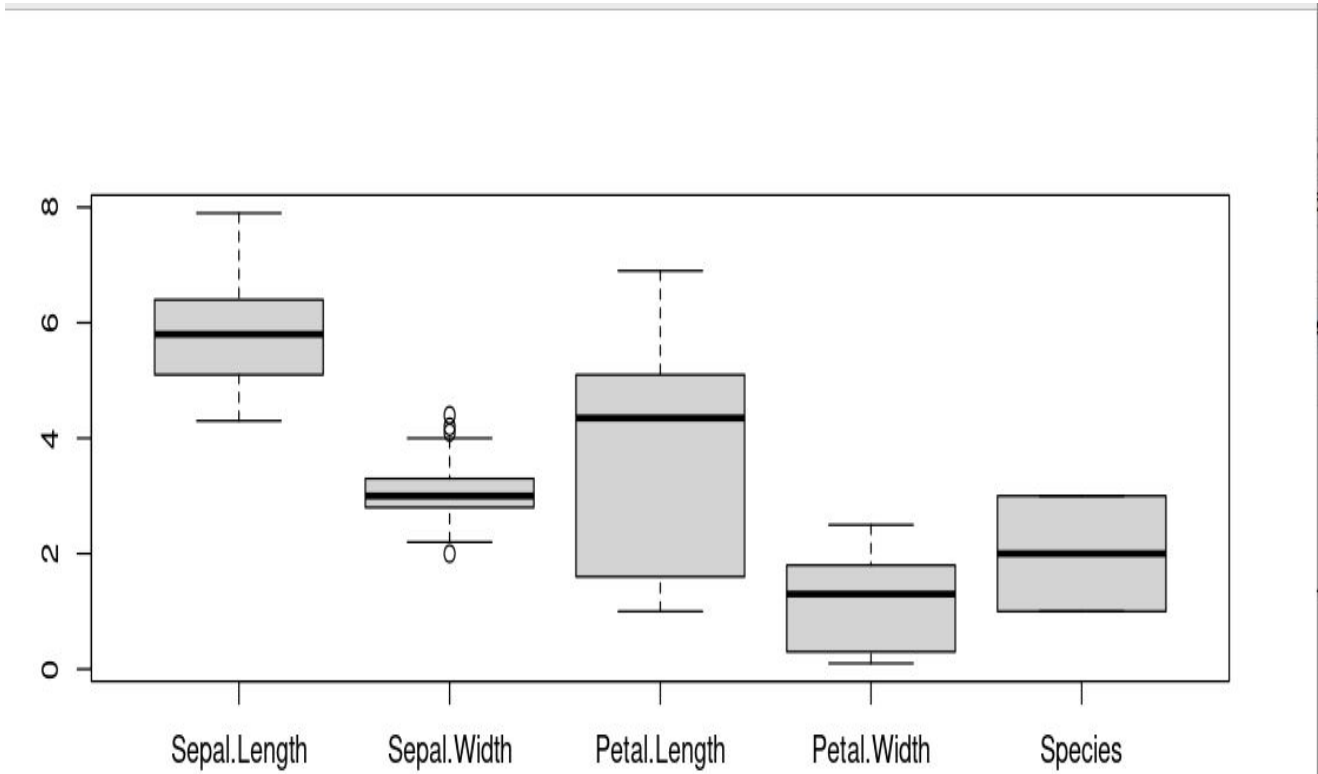
```
> str(dt.iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
> summary(dt.iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

1. Introduction of the dataset

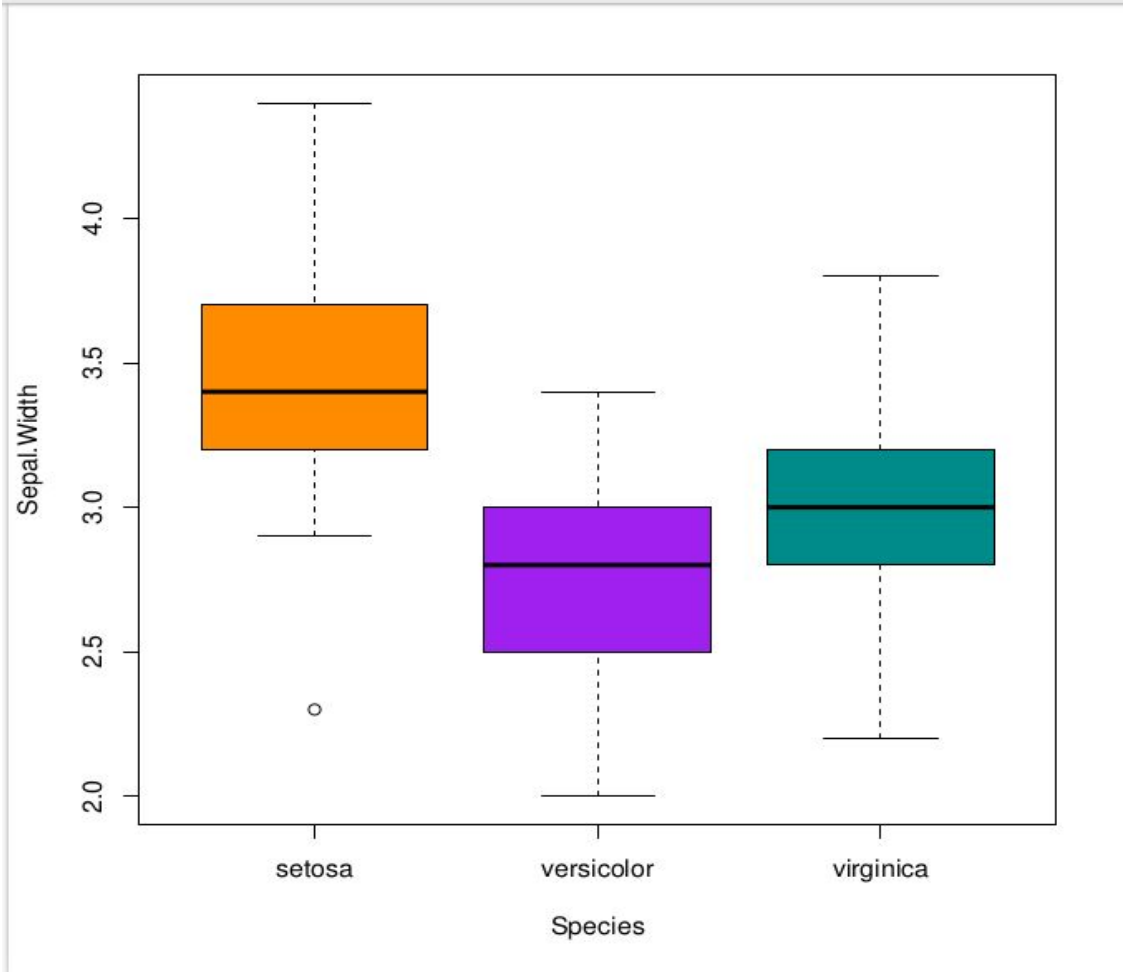
Iris dataset

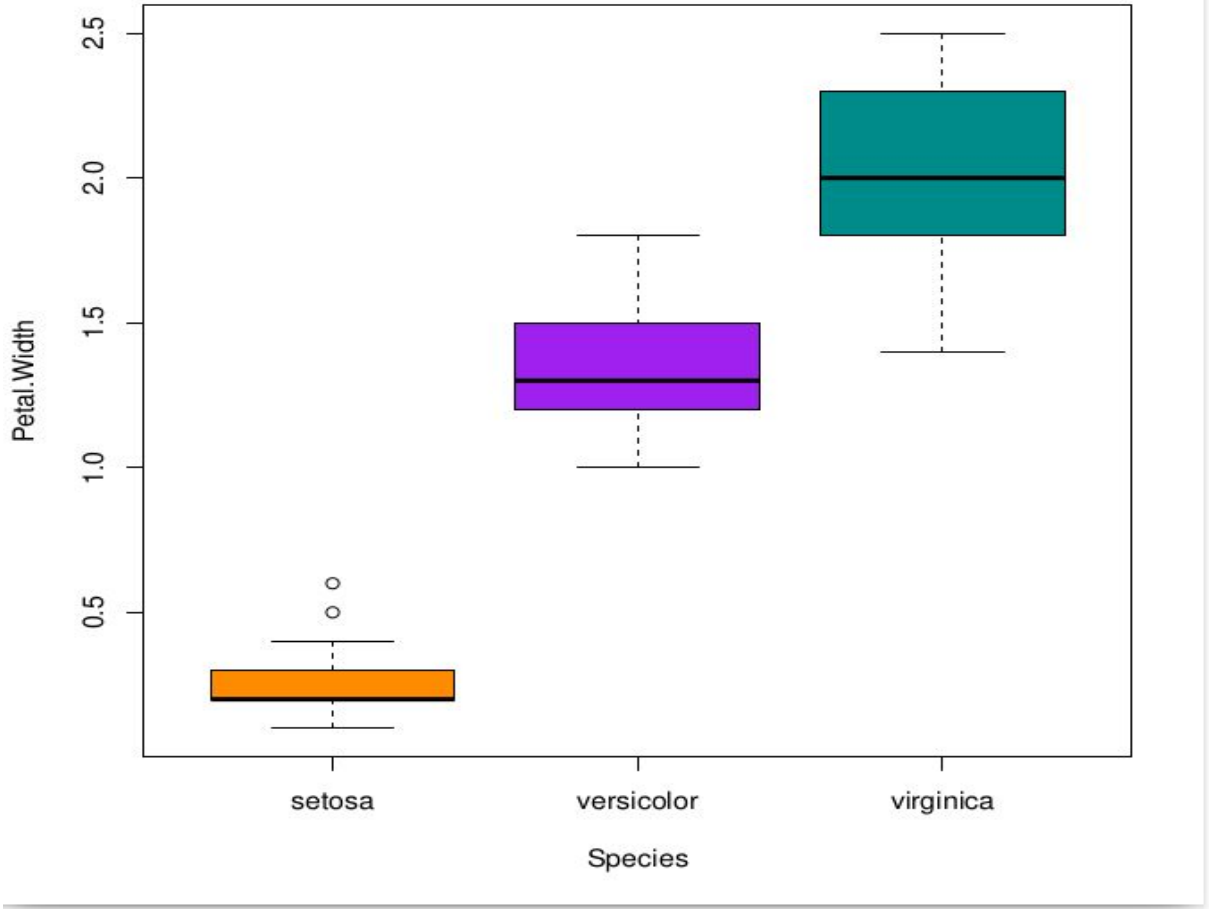




1.Introduction of the dataset

Iris dataset





1.Introduction of the dataset

Microbiome dataset

Murine intestinal microbiome data (Jin et al. 2015) are generated from fecal and cecal stool of vitamin D receptor knockout (Vdr) and wild-type (WT) mice with 454 pyrosequencing. The whole data sets include 5 samples of Vdr mice and 3 samples of WT mice from both fecal and cecal locations. The overall purpose of this study is to explore whether VDR status regulates the composition and functions of the intestinal bacterial community at the genus level.

1.Introduction of the dataset

Microbiome dataset

```
> str(abund_table)
 int [1:16, 1:248] 476 67 549 578 996 404 319 526 424 0 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:16] "5_15_drySt-28F" "20_12_CeSt-28F" "1_11_drySt-28F" "2_12_drySt-28F" ...
  ..$ : chr [1:248] "Tannerella" "Lactococcus" "Lactobacillus" "Lactobacillus::Lactococcus" ...
```

	Tannerella	Lactococcus	Lactobacillus	Lactobacillus::Lactococcus	Parasutterella	Helicobacter	Prevotella	Bacteroides	Barnesiella
5_15_drySt-28F	476	326	94	1	1	89	121	273	9
20_12_CeSt-28F	67	737	597	12	0	0	7	34	1
1_11_drySt-28F	549	2297	434	25	1	0	289	958	2
2_12_drySt-28F	578	548	719	5	4	13	99	377	2
3_13_drySt-28F	996	2378	322	17	2	24	335	526	1
4_14_drySt-28F	404	471	205	1	0	32	143	200	2
7_22_drySt-28F	319	882	644	13	0	3	111	86	0
8_23_drySt-28F	526	1973	2340	15	12	0	89	424	3
9_24_drySt-28F	424	2308	1000	14	1	0	84	202	1
19_11_CeSt-28F	0	422	330	7	0	0	0	0	0
21_13_CeSt-28F	6	173	639	0	0	0	1	5	0
22_14_CeSt-28F	20	580	633	3	0	0	3	20	0
23_15_CeSt-28F	37	4867	1819	25	2	0	10	31	1
25_22_CeSt-28F	38	707	625	9	0	0	7	19	0
26_23_CeSt-28F	81	1404	1361	10	2	0	5	46	0
27_24_CeSt-28F	235	1913	365	13	0	0	25	122	1

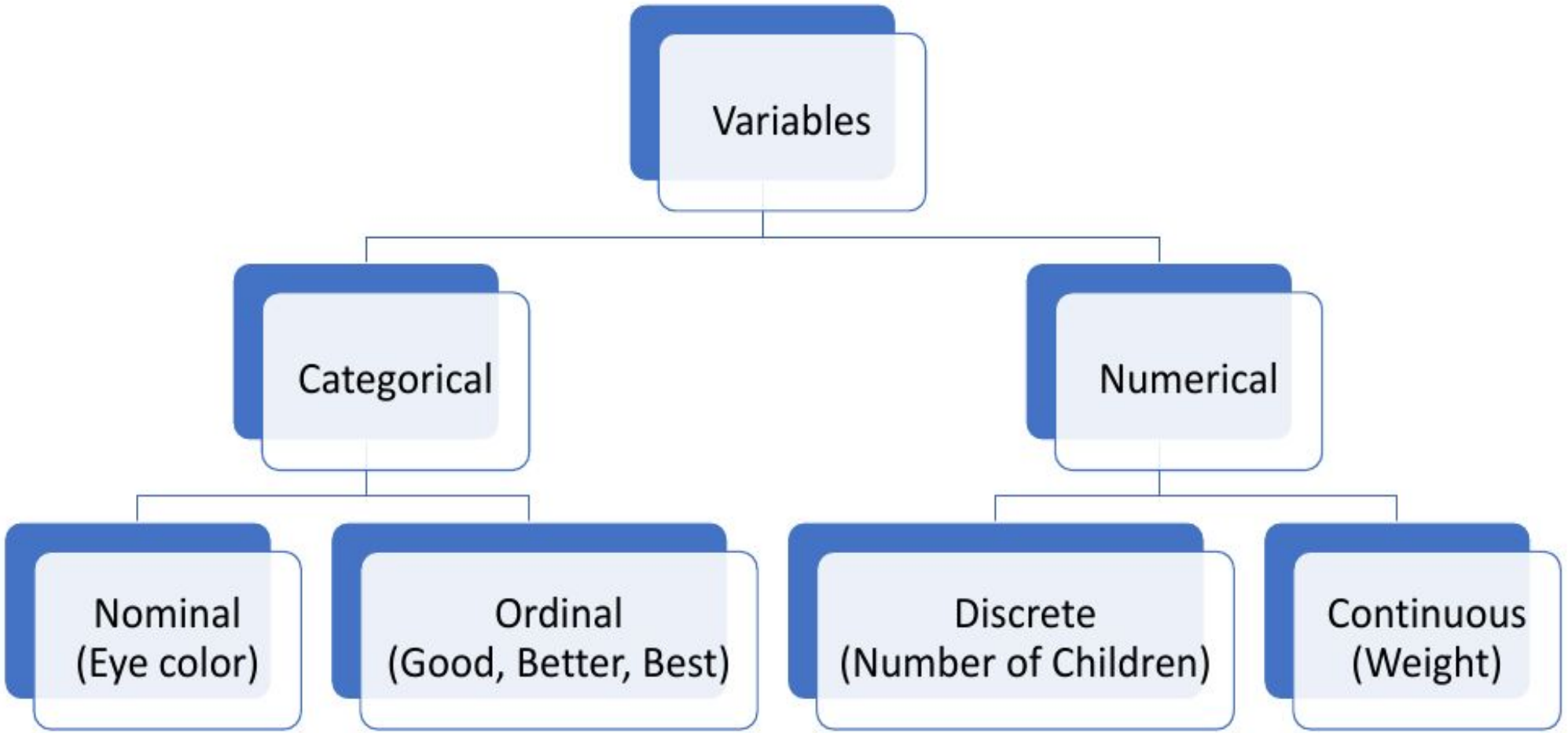
1.Introduction of the dataset

Microbiome dataset

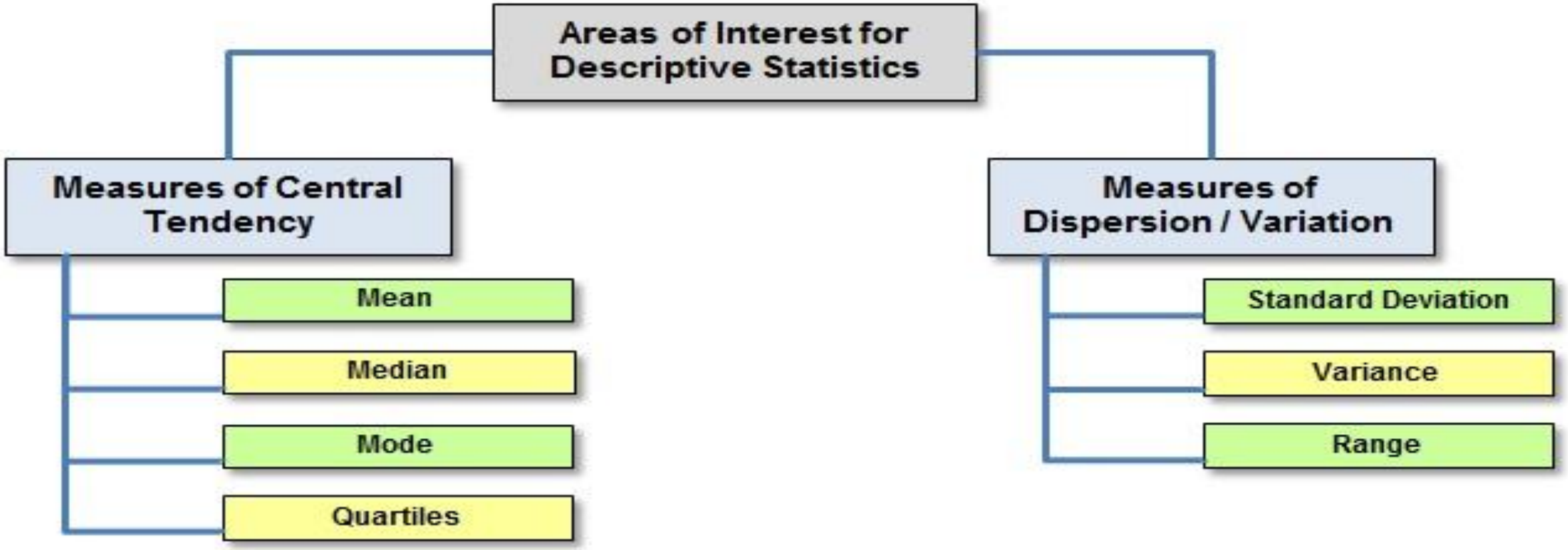
```
>   grouping
                  Location  Group
5_15_drySt-28F       Fecal Vdr-/-
20_12_CeSt-28F       Cecal Vdr-/-
1_11_drySt-28F       Fecal Vdr-/-
2_12_drySt-28F       Fecal Vdr-/-
3_13_drySt-28F       Fecal Vdr-/-
4_14_drySt-28F       Fecal Vdr-/-
7_22_drySt-28F       Fecal      WT
8_23_drySt-28F       Fecal      WT
9_24_drySt-28F       Fecal      WT
19_11_CeSt-28F       Cecal Vdr-/-
21_13_CeSt-28F       Cecal Vdr-/-
22_14_CeSt-28F       Cecal Vdr-/-
23_15_CeSt-28F       Cecal Vdr-/-
25_22_CeSt-28F       Cecal      WT
26_23_CeSt-28F       Cecal      WT
27_24_CeSt-28F       Cecal      WT
```

```
>   table(grouping)
          Group
Location Vdr-/- WT
   Cecal      5  3
   Fecal      5  3
```

2. Types of Variables



3. Descriptive statistics

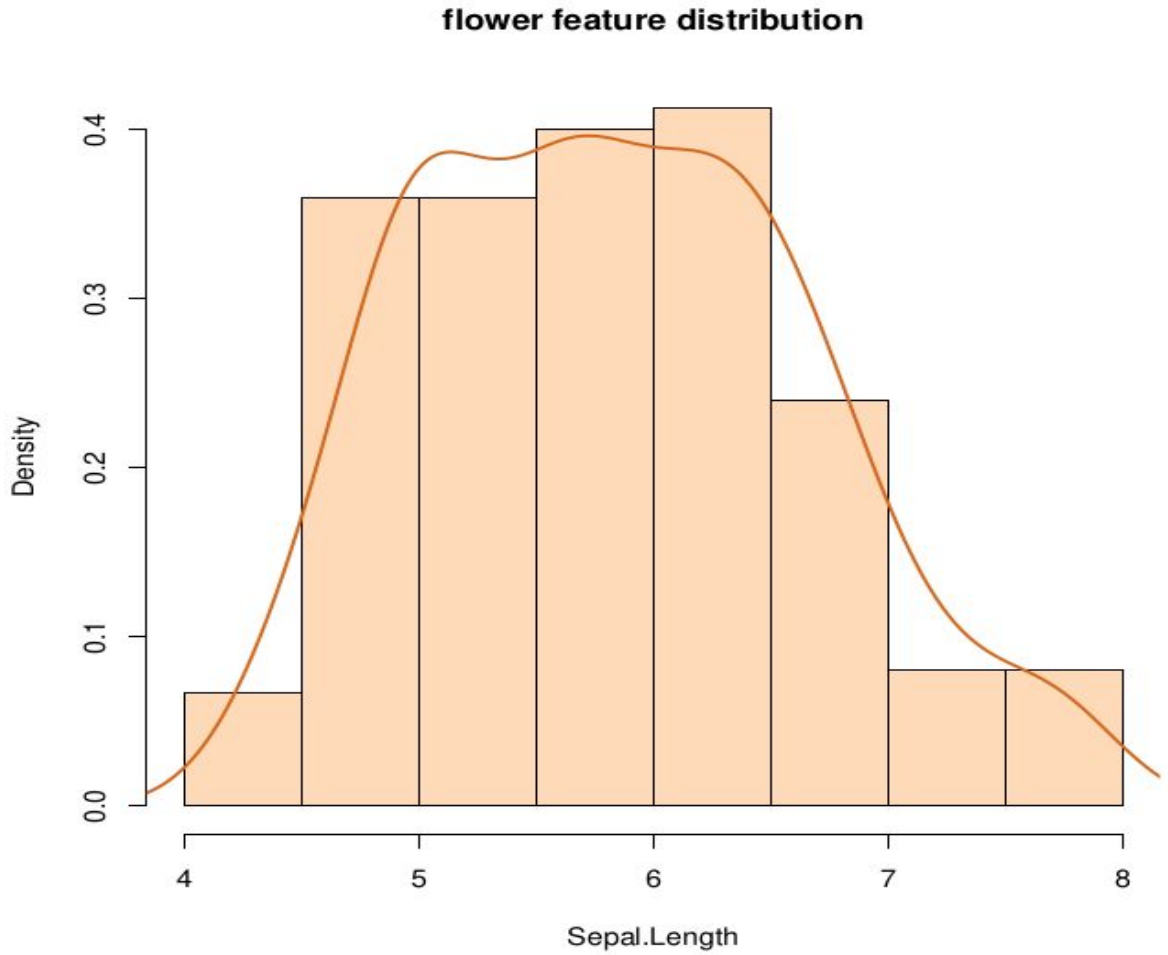


3. Descriptive statistics

MEAN

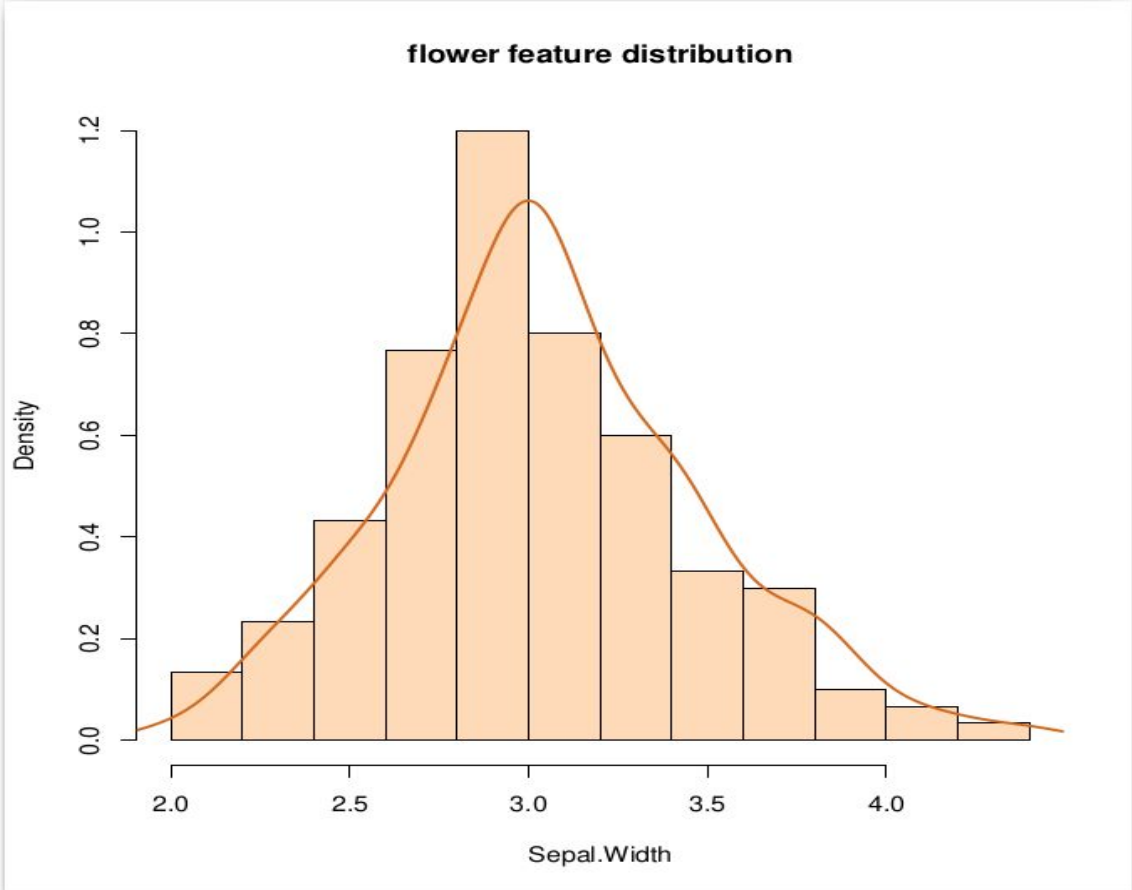
$$\mu_x = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

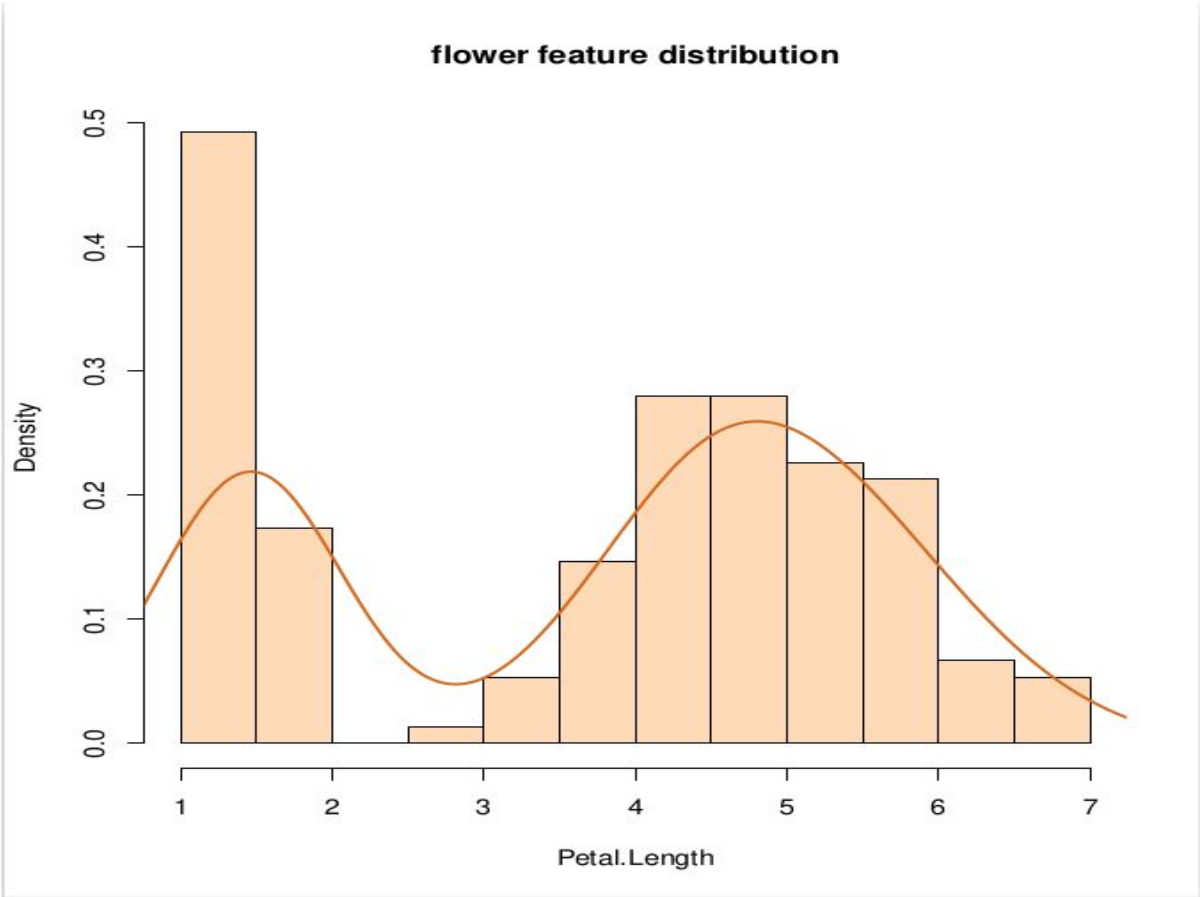
```
> mean(dt.iris$Sepal.Length)
[1] 5.843333
>   mean(dt.iris$Petal.Length)
[1] 3.758
```



3. Descriptive statistics

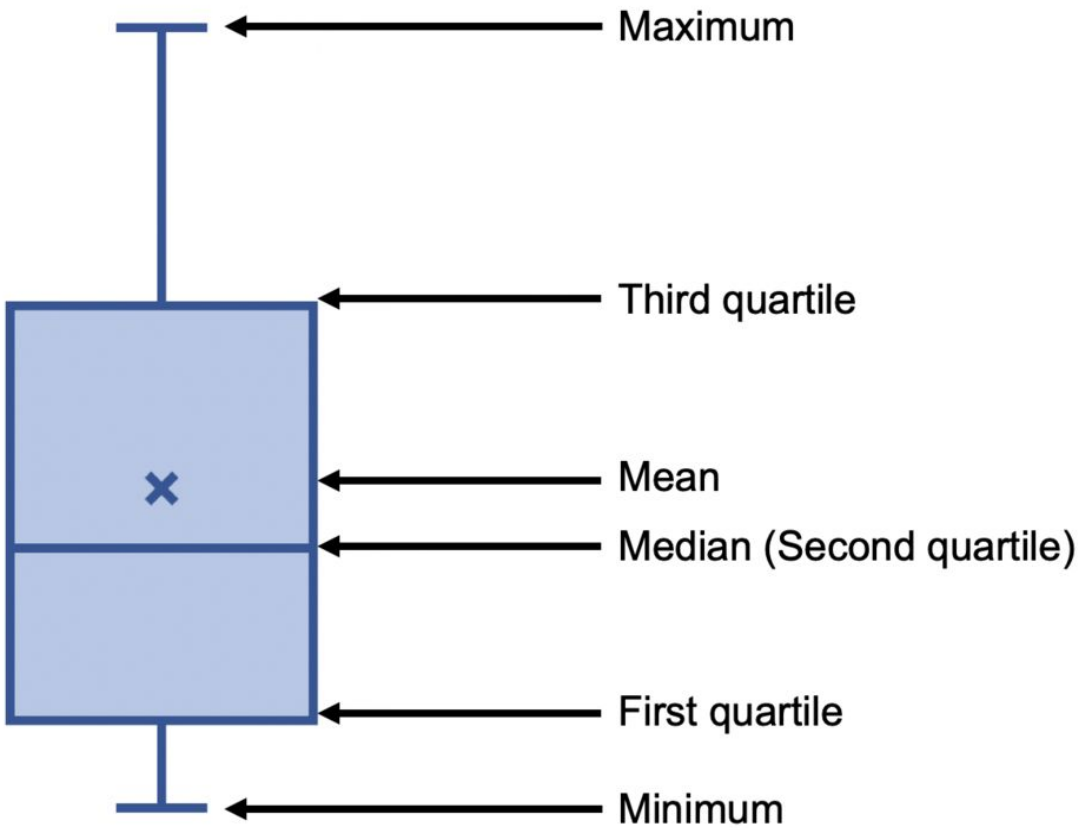
MEAN

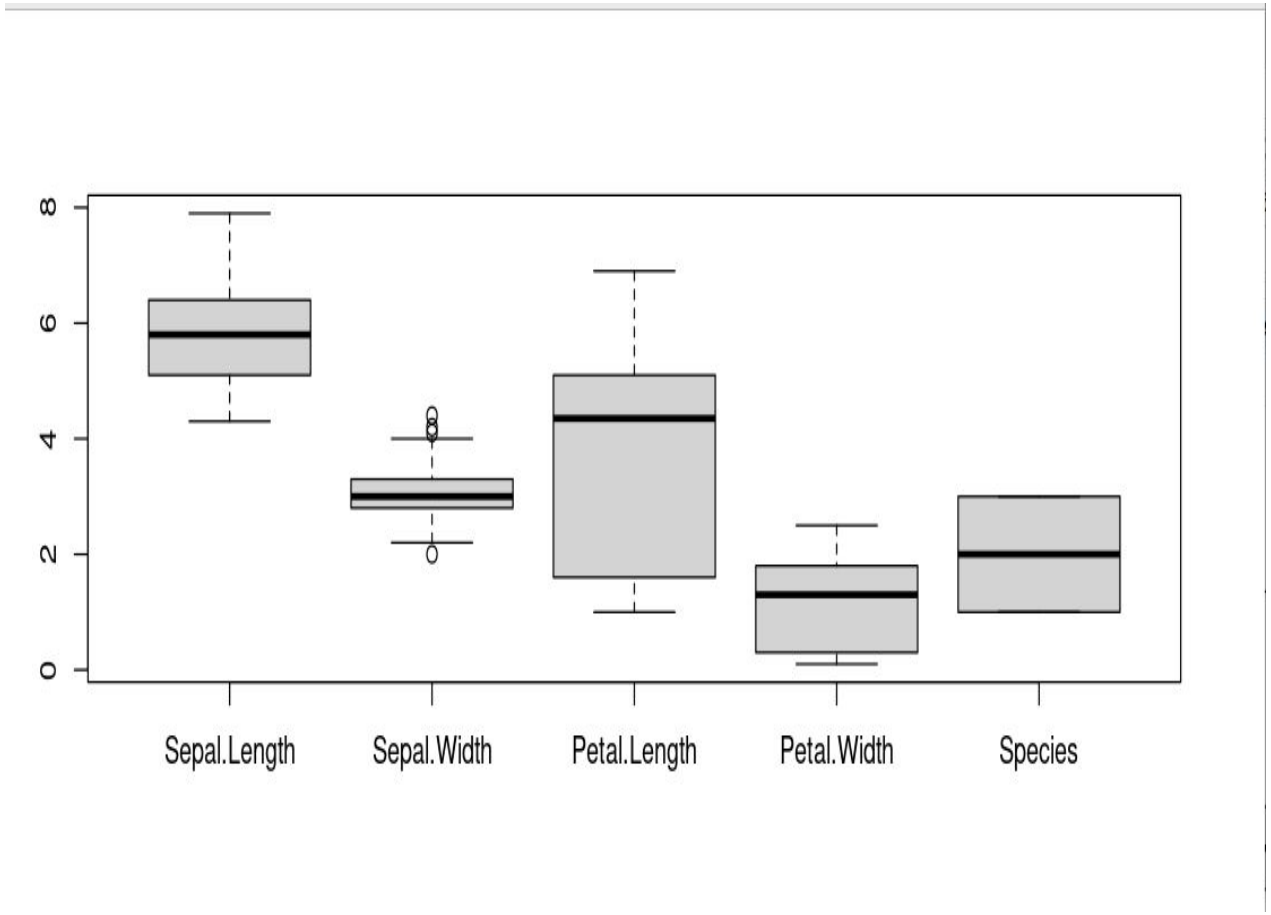




3. Descriptive statistics

MEDIAN and QUARTILES





3. Descriptive statistics

VARIANCE

Measure the dispersion of a set of data values.

$$S_x^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

```
> tapply(dt.iris$Petal.Length, dt.iris$Species, function(x) {
+     variances <- var(x, na.rm=T)
+     return(variances)
+   })
    setosa versicolor  virginica
0.03015918 0.22081633 0.30458776
```

```
> var_Petal_Length
[1] 3.116278
```

3. Descriptive statistics

COVARIANCE

A measure of the joint variability of two variables

- Covariance formula for population:

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

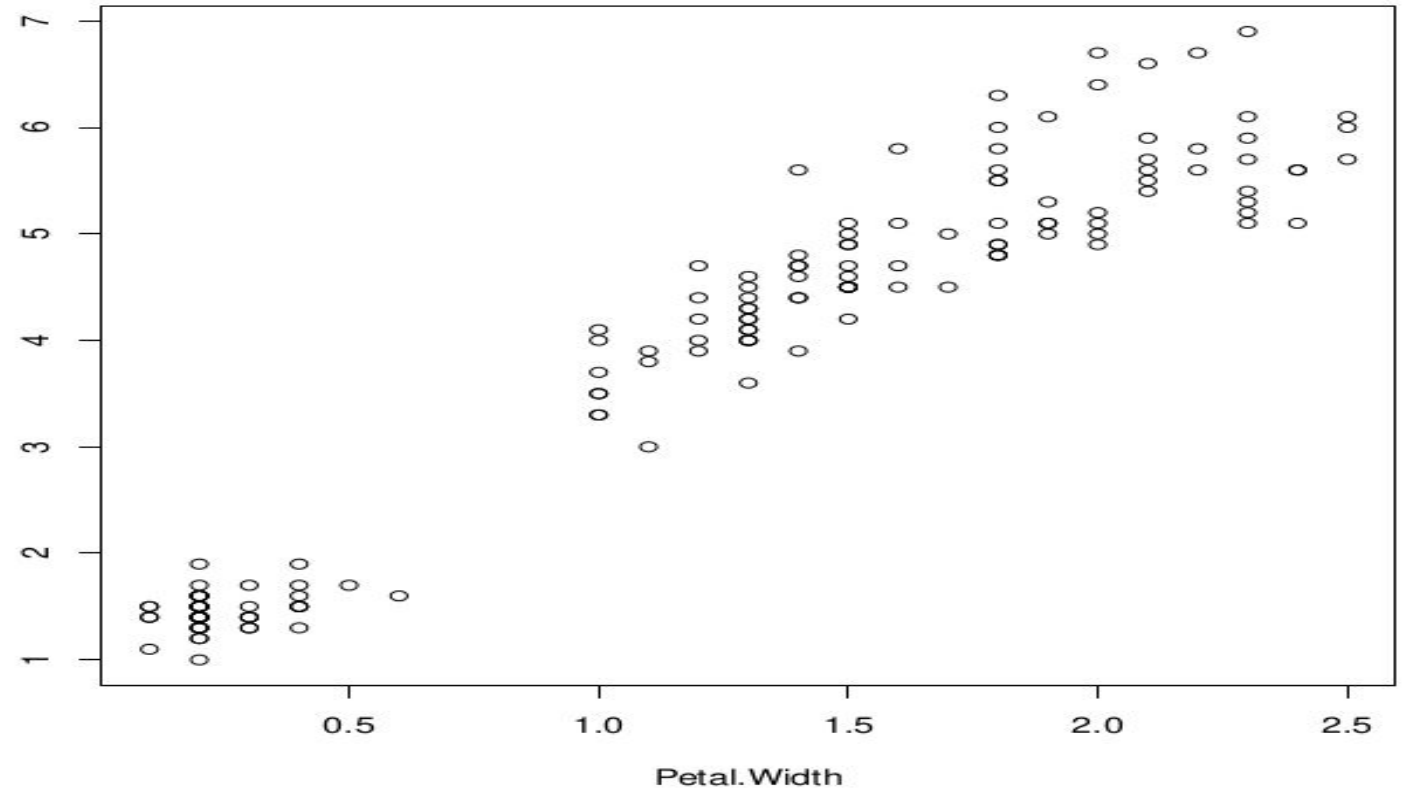
- Covariance Formula for a sample:

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

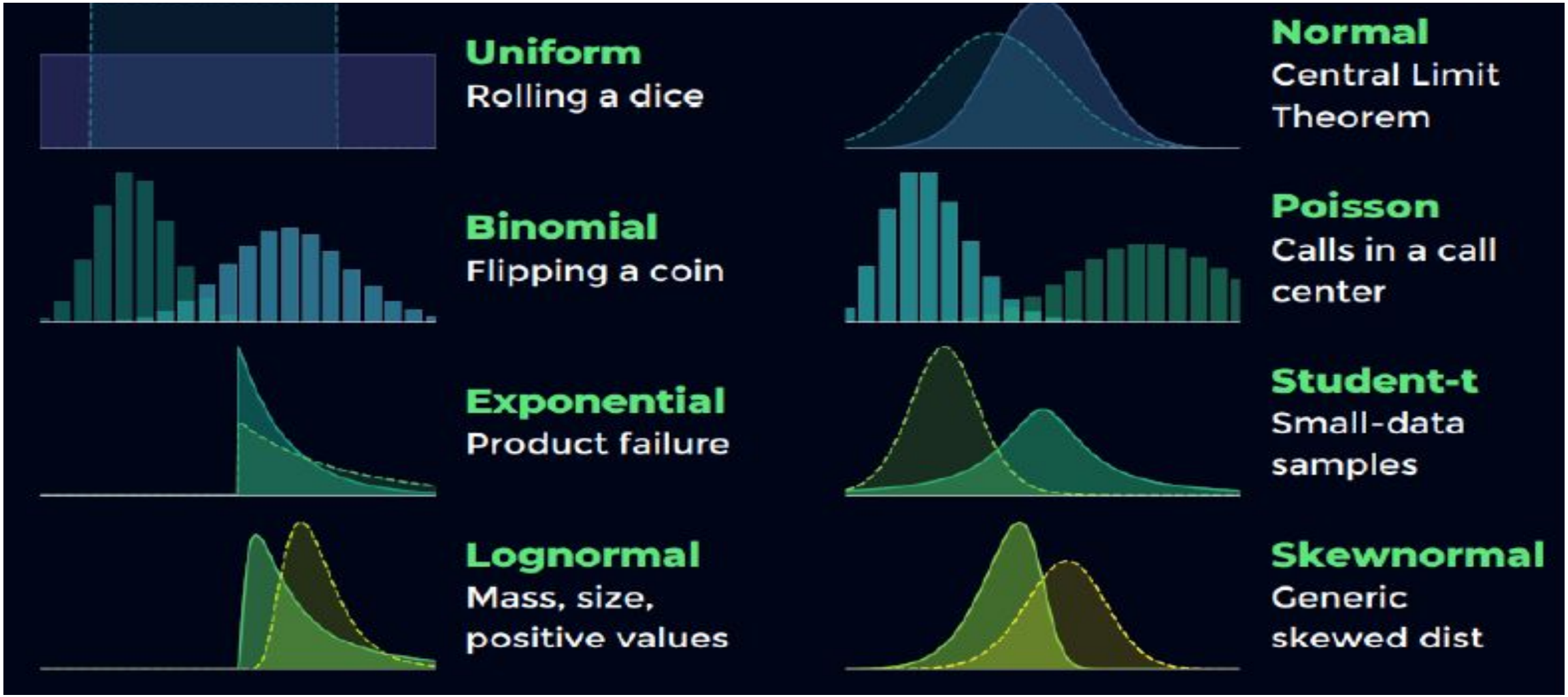
Where,

- X_i is the values of the X-variable
- Y_i is the values of the Y-variable
- $\bar{X}$ is the mean of the X-variable
- $\bar{Y}$ is the mean of the Y-variable
- n is the number of data points

```
> cov_Petal_Width_Length <- cov(dt.iris$Petal.Length,
+                               dt.iris$Petal.Width,
+                               use="complete.obs")
>   cov_Petal_Width_Length
[1] 1.295609
>  var_Petal_Width_Length <- var(dt.iris$Petal.Length,
+                                 dt.iris$Petal.Width,
+                               na.rm=T)
>   var_Petal_Width_Length
[1] 1.295609
```

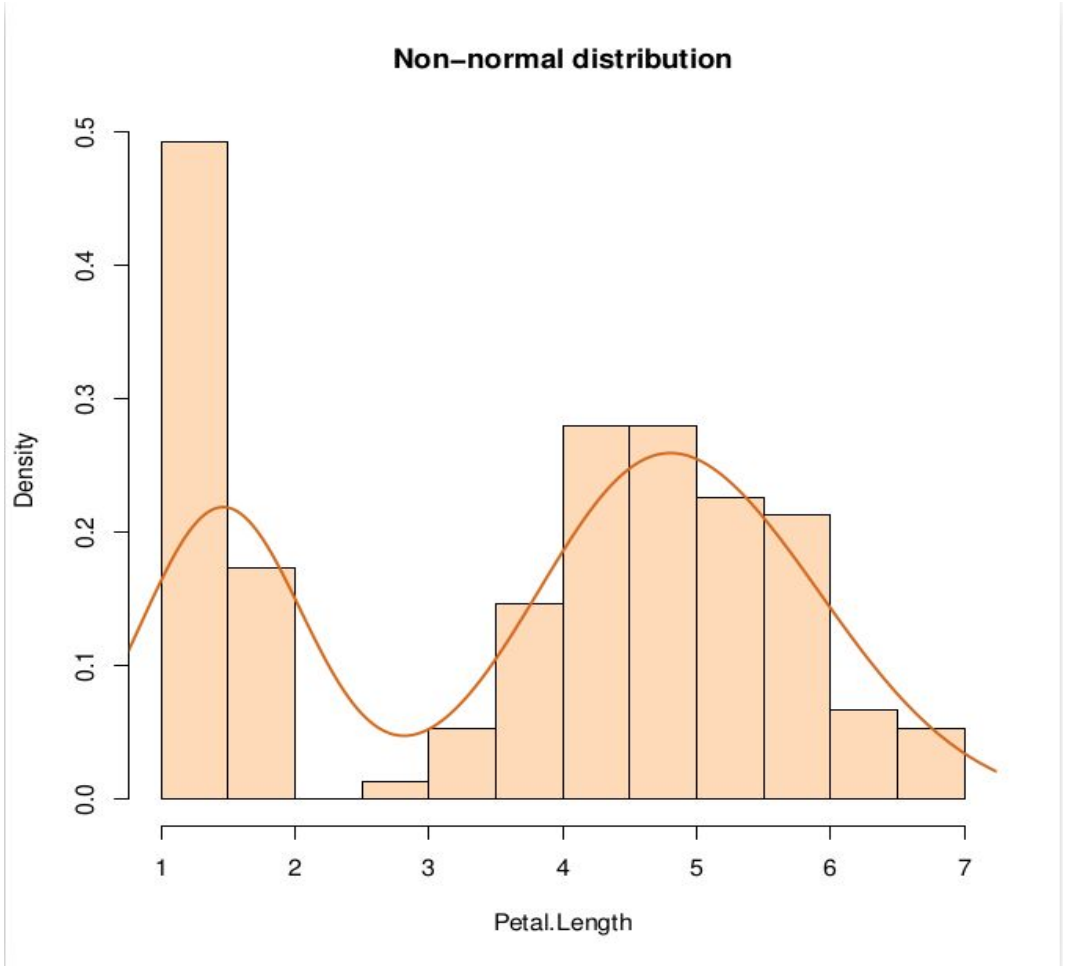


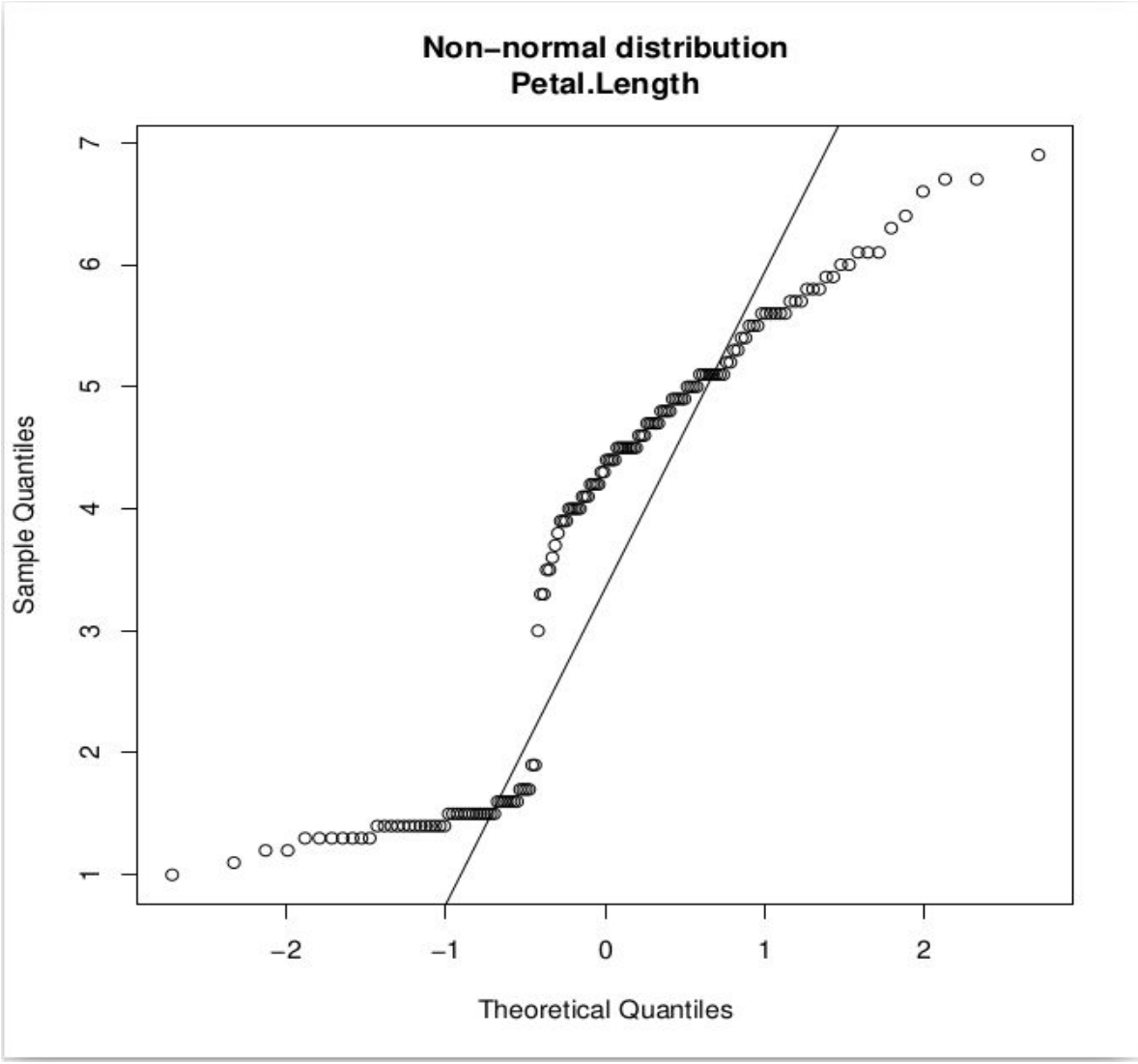
4. Data distribution



4. Data distribution

Check normality of Petal Length in Iris dataset.





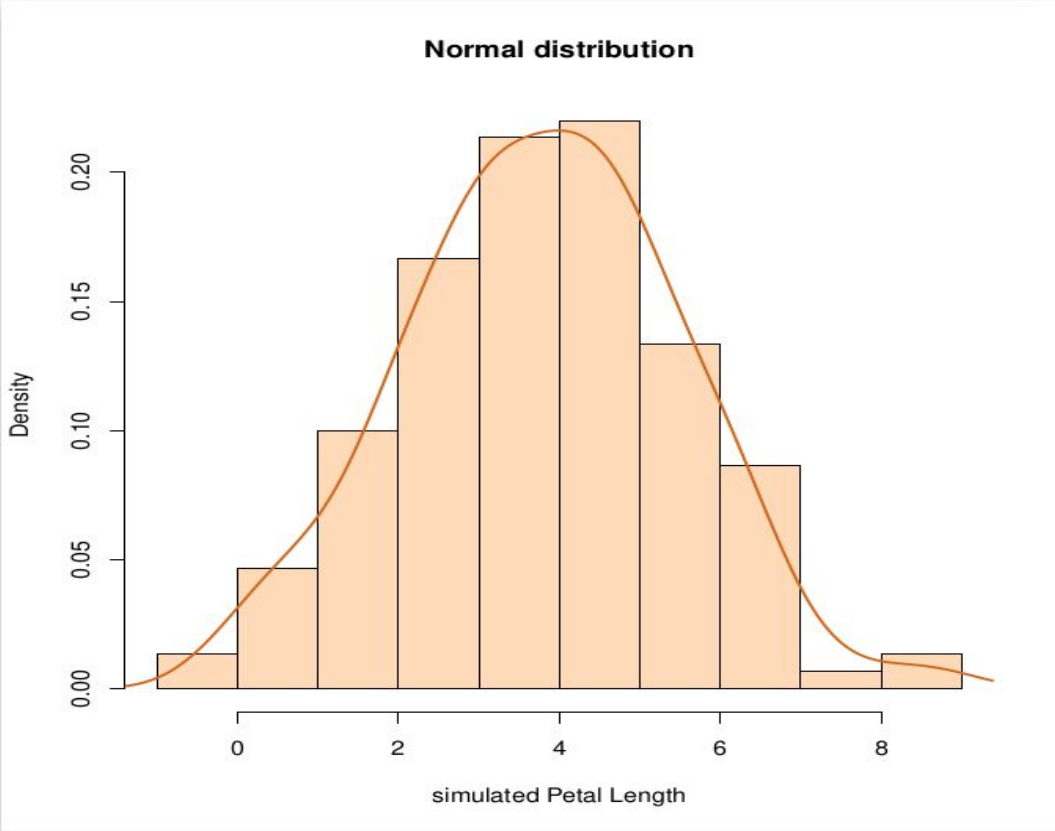
4. Data distribution

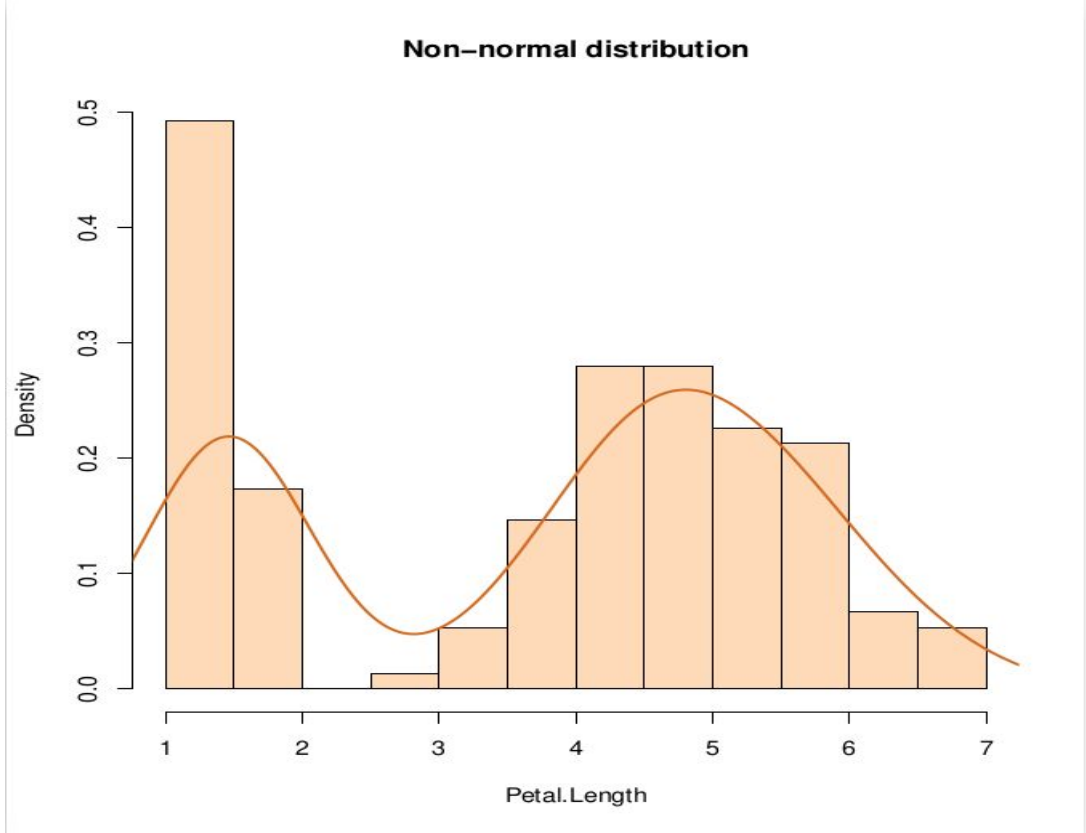
Simulate a normal distributed Petal.Length

```
>   dt.iris$sim_Petal <- NA
>   mean_Petal_Length <- mean(dt.iris$Petal.Length, na.rm=T)
>   sd_Petal_Length <- sd(dt.iris$Petal.Length, na.rm=T)
>   p_values_sw <- rep(0, 9)
>   p_values_ks <- rep(0, 9)
>   iteration <- 0
> while (any(p_values_sw < 0.8) | any(p_values_ks < 0.8)) {
+     iteration <- iteration + 1
+     cat(paste0("Number of iterations: ", iteration, "\n"))
+     dt.iris$sim_Petal <- rnorm(nrow(dt.iris),
+                                mean = mean_Petal_Length,
+                                sd = sd_Petal_Length)
+
+     p_values_sw <- sapply(c("setosa", "versicolor", "virginica"), function(y) {
+         shapiro.test( dt.iris[!is.na(dt.iris$sim_Petal) & (dt.iris$Species == y), "sim_Petal"])$p.value
+       })
+
+
+     p_values_ks <- sapply(c("setosa", "versicolor", "virginica"), function(y) {
+         ks.test(dt.iris[!is.na(dt.iris$sim_Petal) & ( dt.iris$Species == y), "sim_Petal"], "pnorm",
+                 mean = mean(dt.iris[dt.iris$Species == y, "sim_Petal"], na.rm = TRUE),
+                 sd = sd(dt.iris[dt.iris$Species == y, "sim_Petal"], na.rm = TRUE))$p.value
+     })
+   }
```

4. Data distribution

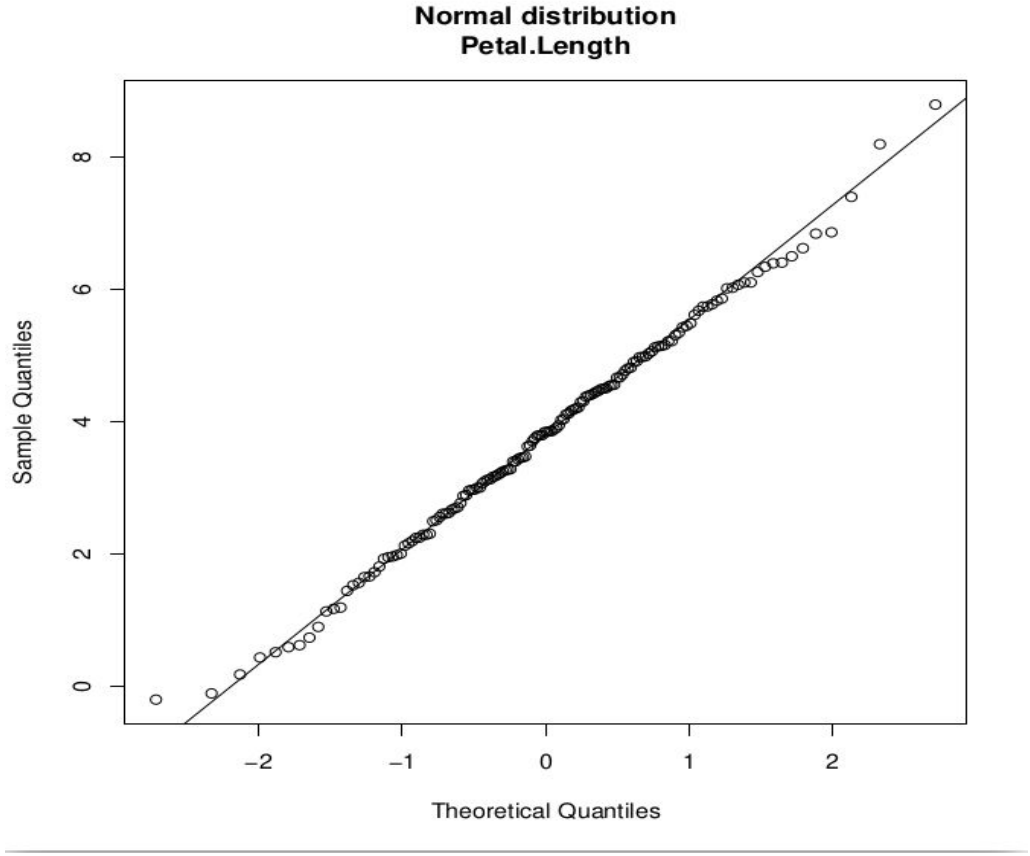
Histogram normality check

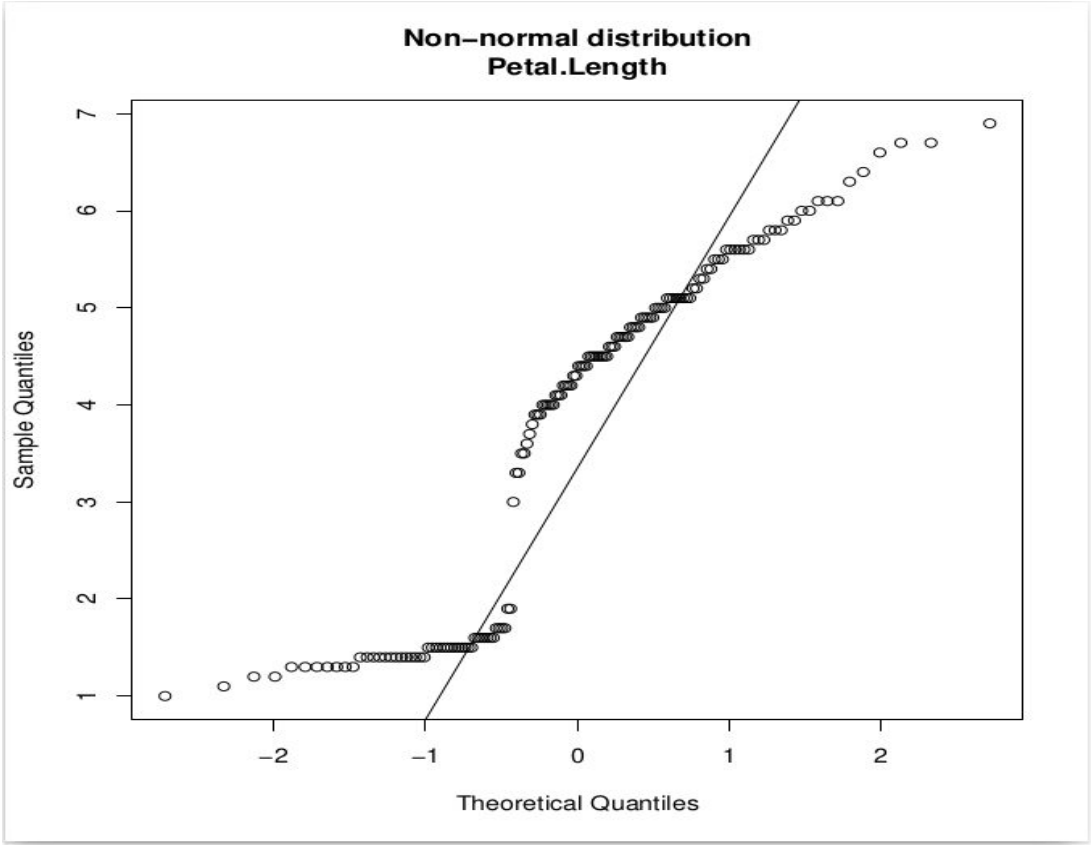




4. Data distribution

Q-Q plot normality check





4. Data distribution

Statistical test normality check

```
>   shapiro.test(dt.iris$Petal.Length)

        Shapiro-Wilk normality test

data:  dt.iris$Petal.Length
W = 0.87627, p-value = 7.412e-10
```

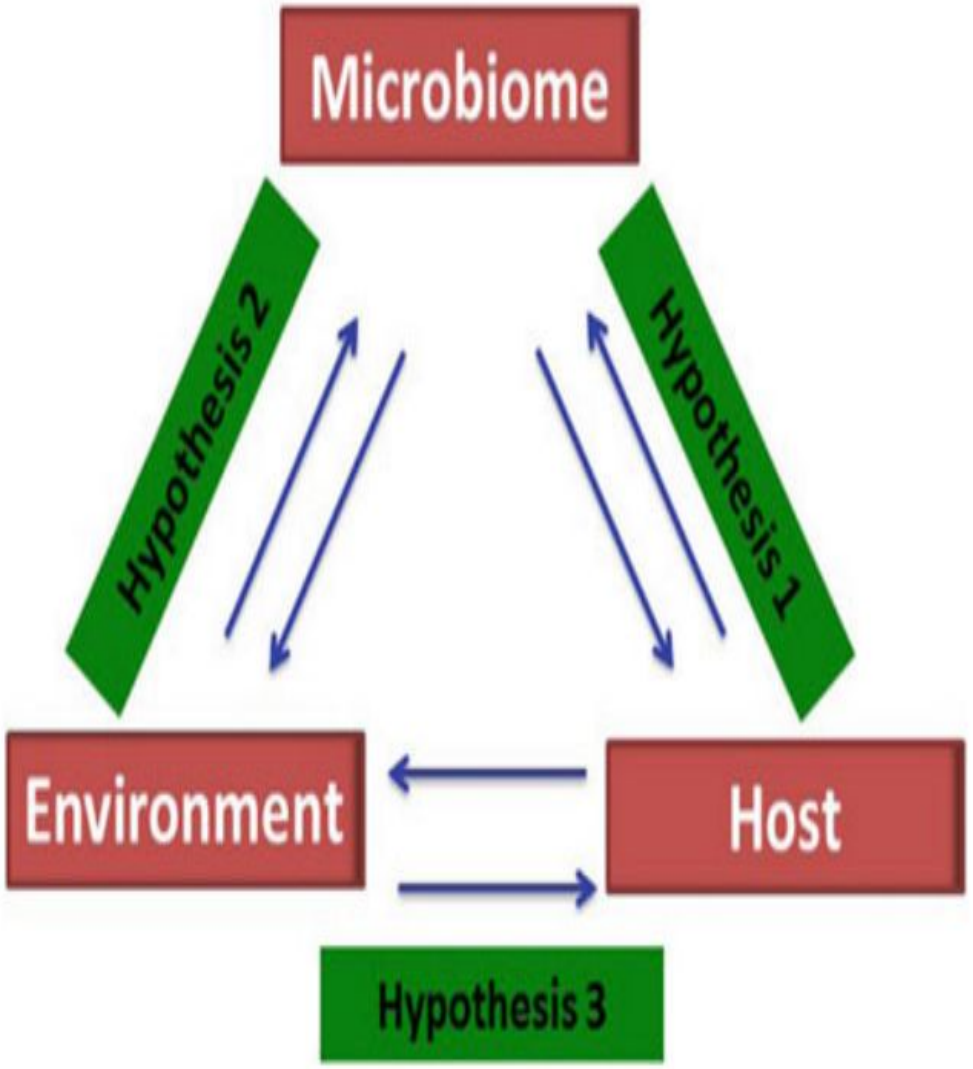
```
>   shapiro.test(dt.iris$sim_Petal)

        Shapiro-Wilk normality test

data:  dt.iris$sim_Petal
W = 0.99593, p-value = 0.9549
```

5. Introduction to hypothesis in microbiome studies

-Ex 1: Hypothesis 1, in inflammatory bowel diseases (IBD) research, hypothesize dysbiosis is associated with the progression of the diseases.

-Ex 2: Hypothesize 2 that antibiotics and diet affect gut microbial community structure, antibiotic treatments affect the diversity of strains of gut bacteria.

Note: For the microbiome studies, the focus is on the hypotheses 1 and 2.



6. Inferential statistics

Parametric test Vs. Non-parametric test

	Parametric test	Non-parametric test
2 Numerical variable (Correlation)	Pearson method	Spearman method
1 Numerical variable 2 Categorical variable (Between groups)	Independent t-test	Wilcoxon Rank Sum test (Mann-Whitney U test)
1 Numerical variable 2 Categorical variable (Within groups)	Paired t-test	Wilcoxon Signed Rank test
1 Numerical variable >2 Categorical variable	One-way ANOVA	Kruskal-Wallis test
2 Categorical variable	-	Chi-squared test

6. Inferential statistics

When will we using parametric tests?

-The data are normally distributed or can be transformed to be normally distributed (Logarithmic, square root, inverse , box-cox transformation and quantile normalization).

-The variance of the population is the same for all groups being compared.

-The data are independent and randomly sampled.

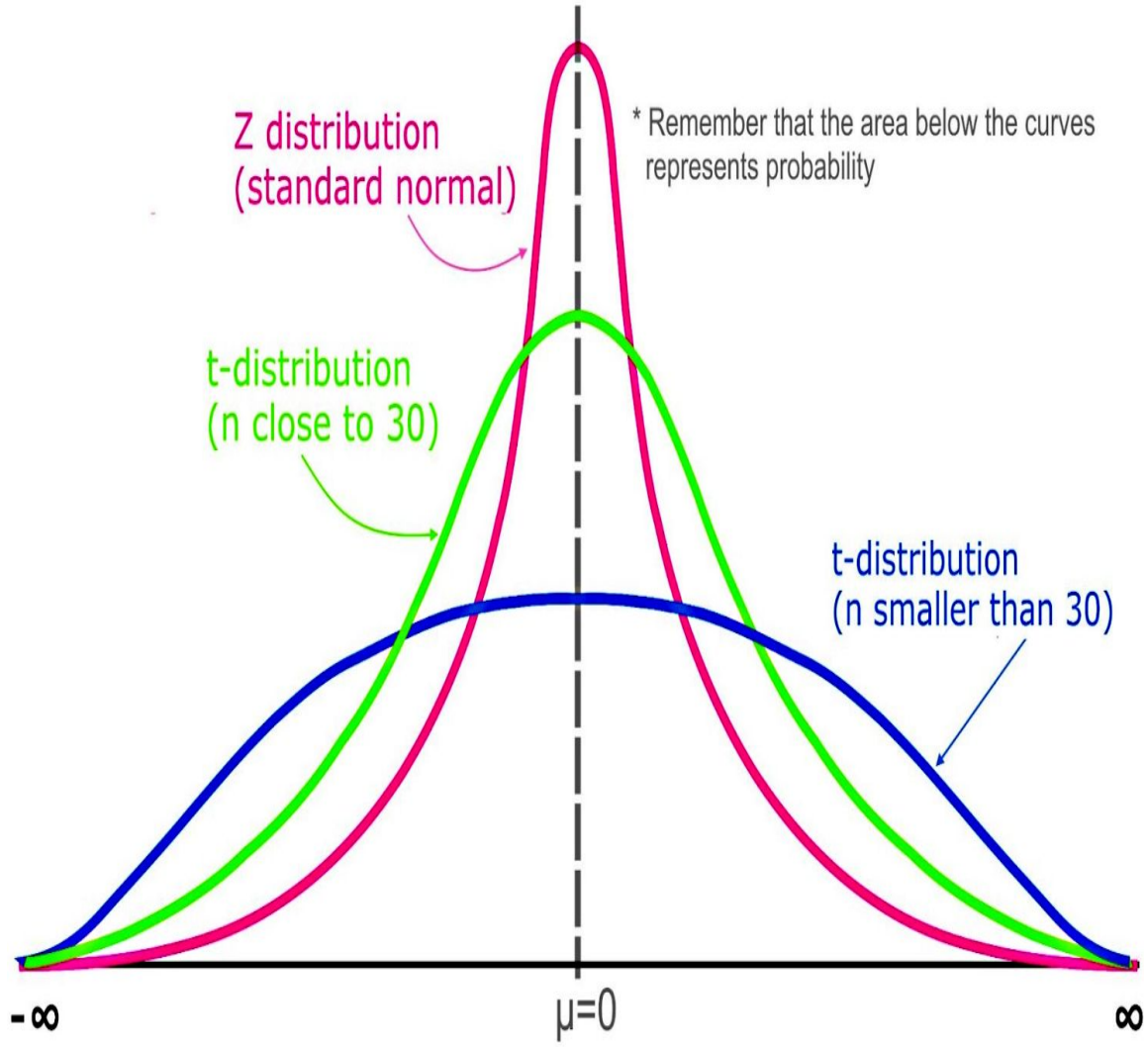
-The dependent variable is continuous or at least ordinal.

7. Hypothesis testing

Two-Sample Welch's t-Test

- Type of parametric statistical test
- Determine whether two groups of data are significantly different from each other
- Based on the t-distribution (probability distribution similar to the normal distribution)

  Assumption: Distribution of two population follow a normal distribution.

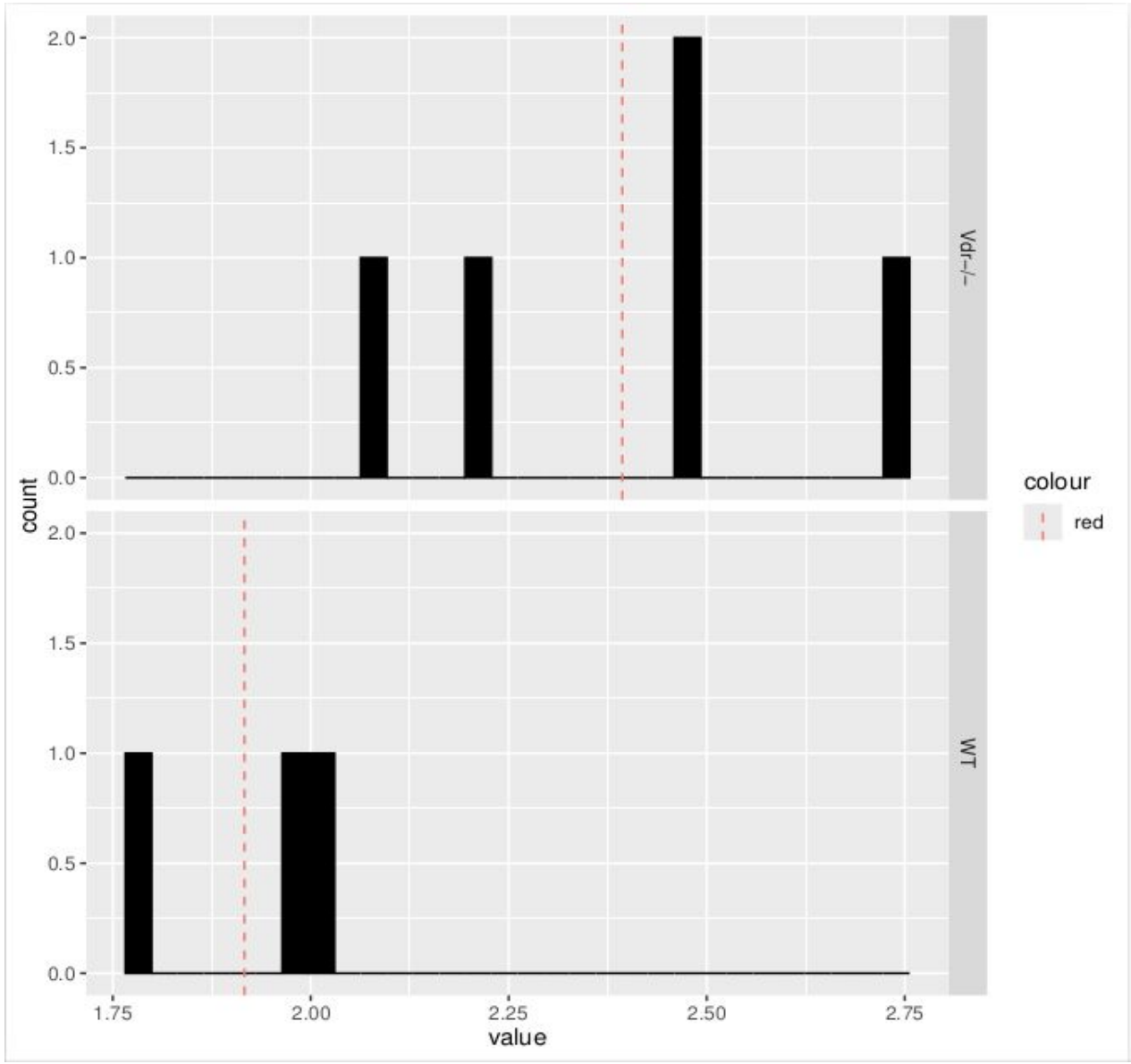


7. Hypothesis testing

Two-Sample Welch's t-Test

```
> library(vegan)
> H<-diversity(abund_table, "shannon")
> df_H<-data.frame(sample=names(H),value=H,measure=rep("Shannon",length(H)))
> df_G <-cbind(df_H, grouping)
```

```
>   Fecal_G
          sample    value measure Location  Group
1 5_15_drySt-28F 2.460729 Shannon    Fecal Vdr-/-
3 1_11_drySt-28F 2.228023 Shannon    Fecal Vdr-/-
4 2_12_drySt-28F 2.734405 Shannon    Fecal Vdr-/-
5 3_13_drySt-28F 2.077282 Shannon    Fecal Vdr-/-
6 4_14_drySt-28F 2.466830 Shannon    Fecal Vdr-/-
7 7_22_drySt-28F 1.777171 Shannon    Fecal     WT
8 8_23_drySt-28F 1.999559 Shannon    Fecal     WT
9 9_24_drySt-28F 1.971996 Shannon    Fecal     WT
```

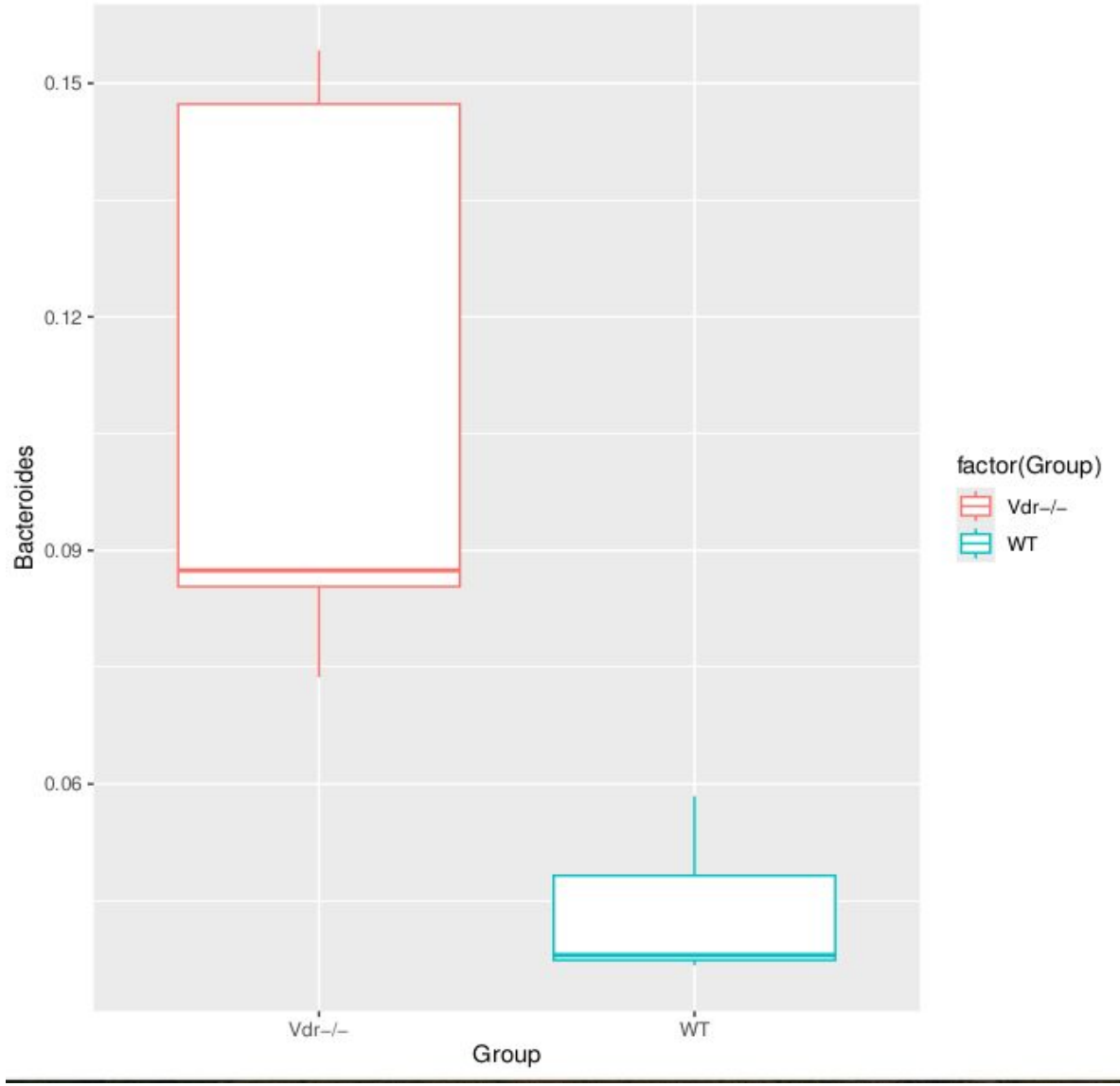


7. Hypothesis testing

Two-Sample Welch's t-Test with microbiome dataset

Note: **Assumption** of t test and **Distribution** of data

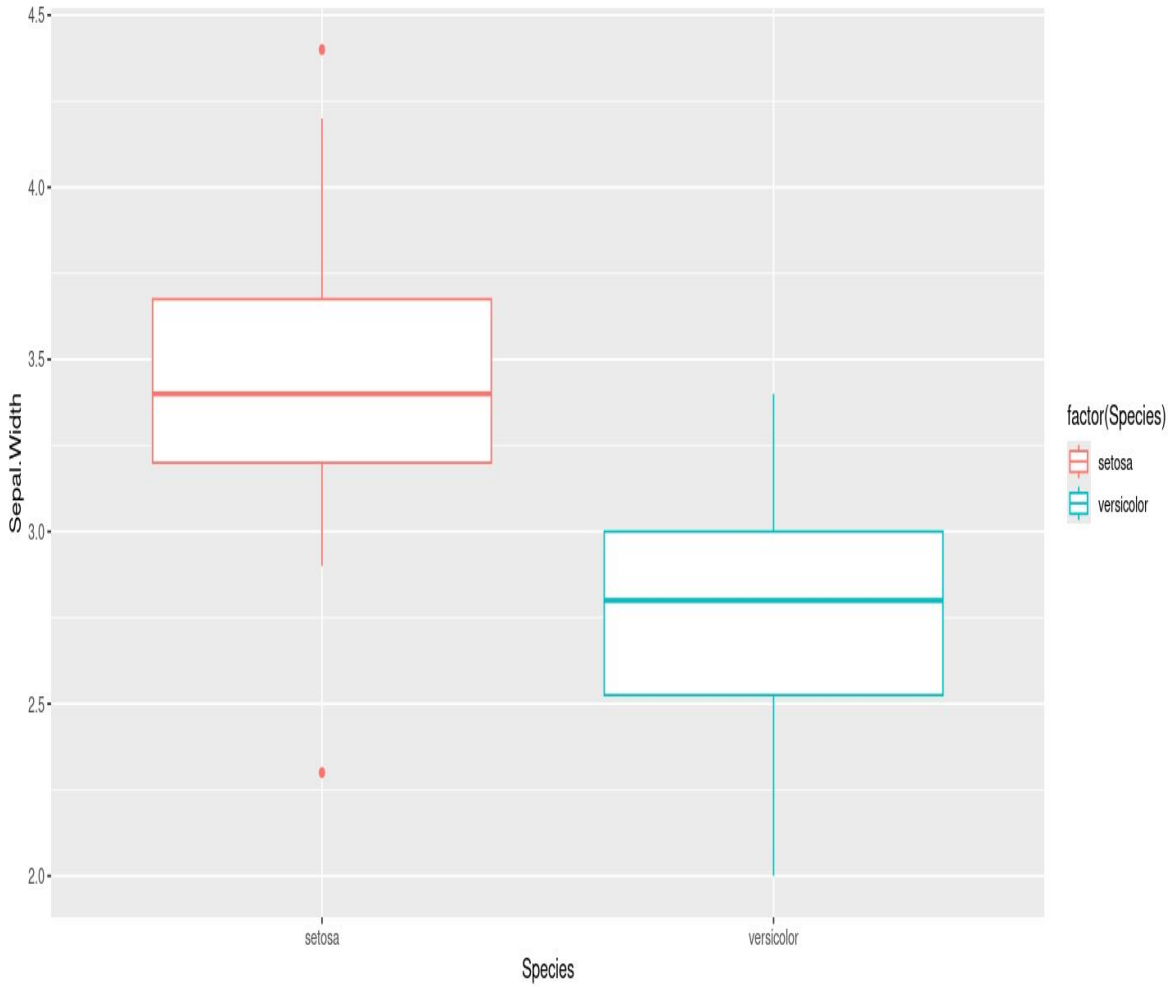
```
> fit_t <- t.test(value ~ Group, data=Fecal_G)
> fit_t

        Welch Two Sample t-test

data:  value by Group
t = 3.5999, df = 5.9206, p-value = 0.01163
alternative hypothesis: true difference in means between group Vdr-/- and group WT is not equal to
0
95 percent confidence interval:
 0.1517841 0.8026392
sample estimates:
mean in group Vdr-/-     mean in group WT
            2.393454            1.916242
```



7. Hypothesis testing

Two-Sample Welch's t-Test with Iris dataset

```
> ttest2 <- t.test(Sepal.Width ~ Species, data= dt.iris_species)
> ttest2

        Welch Two Sample t-test

data:  Sepal.Width by Species
t = 9.455, df = 94.698, p-value = 2.484e-15
alternative hypothesis: true difference in means between group setosa and group versicol
or is not equal to 0
95 percent confidence interval:
 0.5198348 0.7961652
sample estimates:
    mean in group setosa mean in group versicolor
                   3.428                    2.770
```



7. Hypothesis testing

Analysis of variance (ANOVA) test

- Type of parametric statistical test

- Compare the mean three or more groups of data

- The null hypothesis of ANOVA is: all the means of compared groups are equal.

Assumption: Normality of the underlying data -> ANOVA is only used for comparing univariate analysis of alpha diversity measures

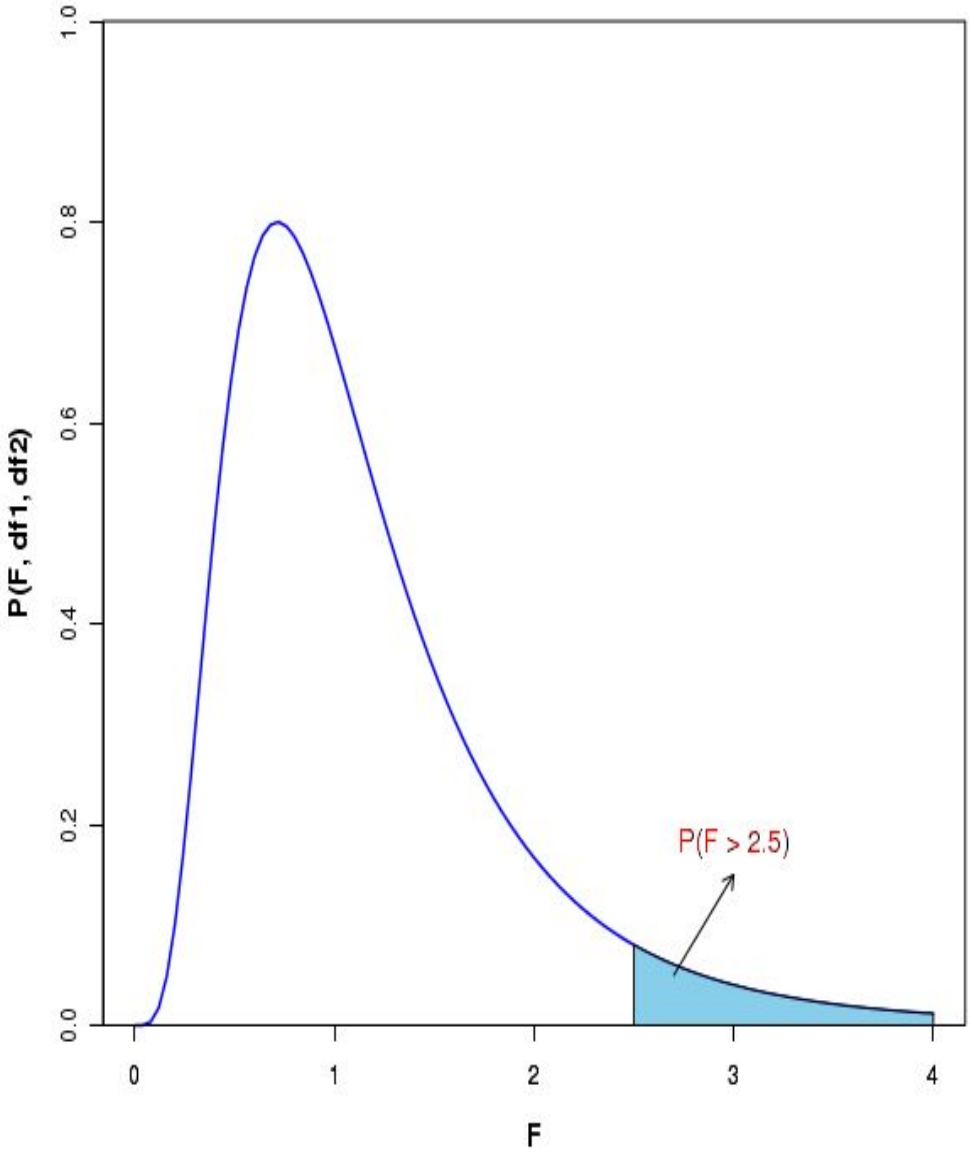
7. Hypothesis testing

ANOVA test

- The formation of testing statistic is through using traditional partitioning of the sum of squares.
- The F-test is used for comparing the factors of the total deviation.

$$SS_{Total} = SS_{Treatments} + SS_{Error}$$

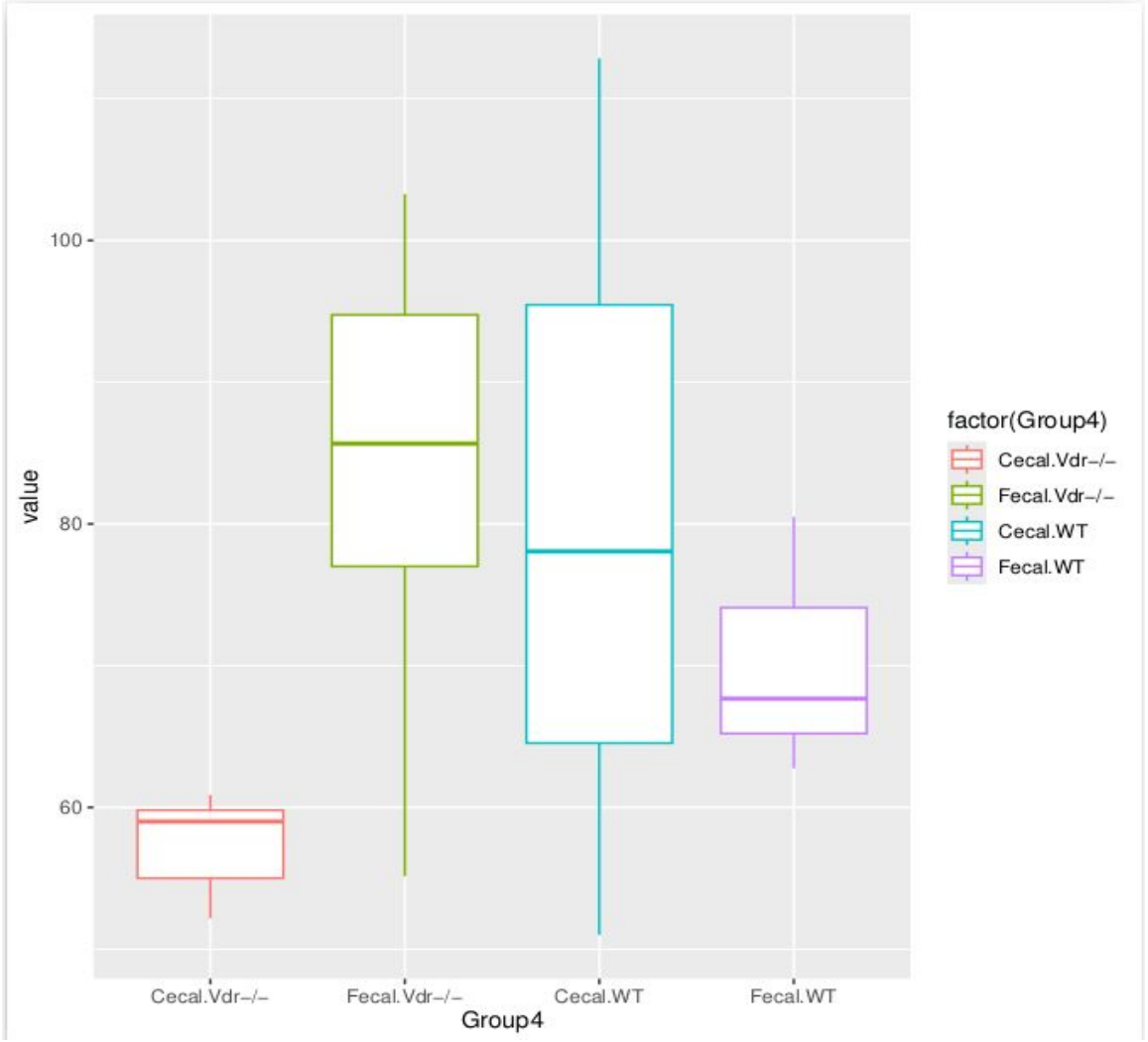
$$F = \frac{MS_{Treatments}}{MS_{Error}} = \frac{SS_{Treatments}/(K-1)}{SS_{Error}/(N-1)}$$



7. Hypothesis testing

ANOVA test

```
>  summary(aov_fit, intercept=T)
             Df Sum Sq Mean Sq F value   Pr(>F)
(Intercept)  1  83450   83450 285.080 9.97e-10 ***
Group4       3   1926     642   2.193    0.142
Residuals   12   3513     293
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>  qf(0.95, 12, 3)
[1] 8.744641
```



7. Hypothesis testing

NOTE: To proceed with the verification of using **ANOVA,** we must first test for homogeneity of variances: the Bartlett's test,

and the Fligner-Killeen test.

To illustrate the test for **homogeneity of variances**, we use the Chao1 richness measures of Vdr and WT mouse data from both fecal and cecal locations.

The null hypothesis (H0) is that **all variances** in four groups are the **same**

Compare the Bartlett's **K-squared** with the value of **chi-square tables**, using the **same level of alpha** and **degrees of freedom**.

7. Hypothesis testing

Bartlett test: Test for homogeneity of variances

```
> bartlett.test(value ~ Group4,  df_CH_G4)

        Bartlett test of homogeneity of variances

data:  value by Group4
Bartlett's K-squared = 10.227, df = 3, p-value = 0.01673

>   qchisq(0.95, 3)
[1] 7.814728
```

7. Hypothesis testing

Fligner-Killeen test: Test for homoscedasticity

```
> df_CH_G4 <- select(df_CH_G, Group4,value)
> fligner.test(df_CH_G4, Group4)

        Fligner-Killeen test of homogeneity of variances

data:  df_CH_G4
Fligner-Killeen:med chi-squared = 20.572, df = 1, p-value = 5.742e-06
```

Chi-square test of independence

Chi-square test is used to determine whether or not there is a significant association between two categorical variables.

Assumption:

- The categories of the variables are **mutually exclusive** (1 subject fits into one and only one cell)
- The study groups must be independent
- Sample size is large enough: the expected frequency in each cell should be at least 5 in at least 80% of the cells

Example 1: A researcher wanted to know if the distribution of Candida is associated with body site

A data.frame: 3 × 3

	absent	present	total
	<dbl>	<dbl>	<dbl>
stool	100	70	170
lung	100	200	300
total	200	270	470

Chi-square test of independence

H0: There is not a relationship between the body site and the present of Candia

Ha: There is a relationship between the body site and the present of Candia

Chi-Square Test Statistic

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Expected Cell Value

$$E = \frac{row\ total \times column\ total}{n}$$

Degree of freedom: (total column - 1) x (total row - 1)

Example 1: A researcher wanted to know if the distribution of Candida is associated with body site

A data.frame: 3 × 3

	absent	present	total
	<dbl>	<dbl>	<dbl>
stool	100	70	170
lung	100	200	300
total	200	270	470

Chi-square test of independence

```r
# Stool : absent - present
stool <- c(100, 70)
# Lung : absent - present
lung <- c(100, 200)

total <- stool + lung
data = as.data.frame(rbind(stool, lung, total))
data$row_total = rowSums(data)

colnames(data) = c('absent', 'present', 'total')
data

chisq.test(data[1:2, 1:2])
```

Example 1: A researcher wanted to know if the distribution of Candida is associated with body site

A data.frame: 3 × 3

	absent	present	total
	<dbl>	<dbl>	<dbl>
stool	100	70	170
lung	100	200	300
total	200	270	470

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

data:  data[1:2, 1:2]
X-squared = 27.808, df = 1, p-value = 1.339e-07
```

Fisher’s exact test

Fisher’s exact test is used to assess the association between two binary variables in a **contingency table**

Assumption:

- The categories of the variables are mutually exclusive (1 subject fits into one and only one cell)
- The study groups must be independent
- **Sample size is small**: more than 20% of the cell are less than 5

Example 1: A researcher wanted to know if the distribution of Aspergillus is associated with body site

	absent	present	total
	<dbl>	<dbl>	<dbl>
stool	10	7	17
lung	4	3	7
total	14	10	24

Fisher’s exact test

H0: There is not a relationship between the body site and the present of Aspergillus

Ha: There is a relationship between the body site and the present of Aspergillus

$$\Pr(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

X = stool/absent
N: sample size
n : total of row 1
K : total of column 1

Example 1: A researcher wanted to know if the distribution of Aspergillus is associated with body site

	absent	present	total
	<dbl>	<dbl>	<dbl>
stool	10	7	17
lung	4	3	7
total	14	10	24

Fisher’s exact test

```r
# Stool : absent - present
stool <- c(10, 7)
# Lung : absent - present
lung <- c(4, 3)

total <- stool + lung
data = as.data.frame(rbind(stool, lung, total))
data$row_total = rowSums(data)

colnames(data) = c('absent', 'present', 'total')
data

chisq.test(data[1:2, 1:2])
fisher.test(data[1:2, 1:2])
```

	absent	present	total
	<dbl>	<dbl>	<dbl>
stool	10	7	17
lung	4	3	7
total	14	10	24

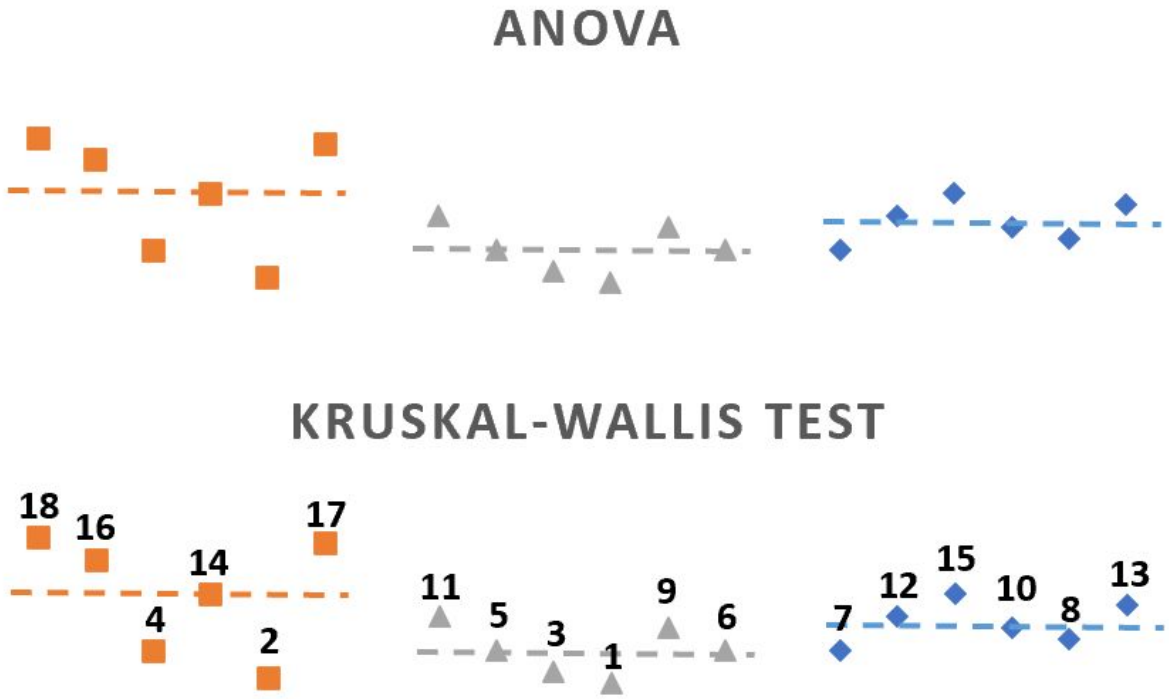
```
Warning message in chisq.test(data[1:2, 1:2]):
“Chi-squared approximation may be incorrect”

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

data:  data[1:2, 1:2]
X-squared = 1.4717e-31, df = 1, p-value = 1

        Fisher's Exact Test for Count Data

data:  data[1:2, 1:2]
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1175924 8.7547429
sample estimates:
odds ratio
  1.068345
```

Kruskal-Wallis-Test

The Kruskal–Wallis test: compare more than **2 groups** for a continuous or discrete variable.

Non-parametric test, it assumes no particular distribution of your data

Analogous to the one-way analysis of variance (ANOVA)

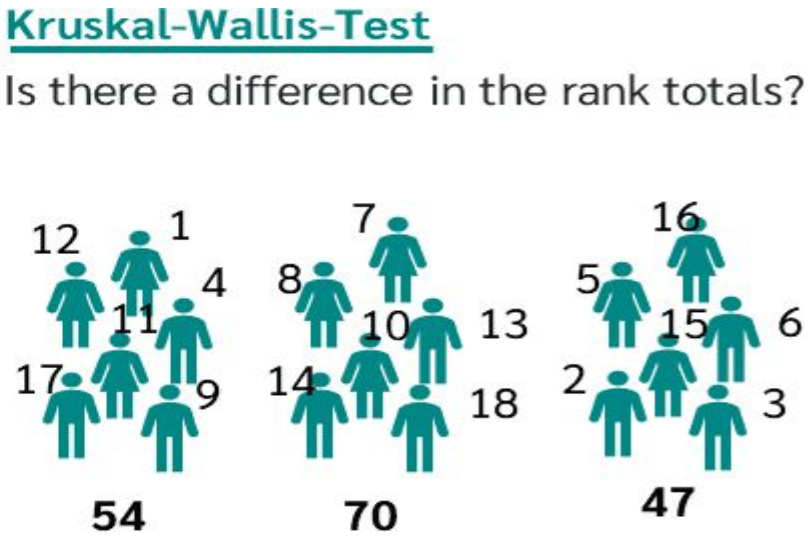


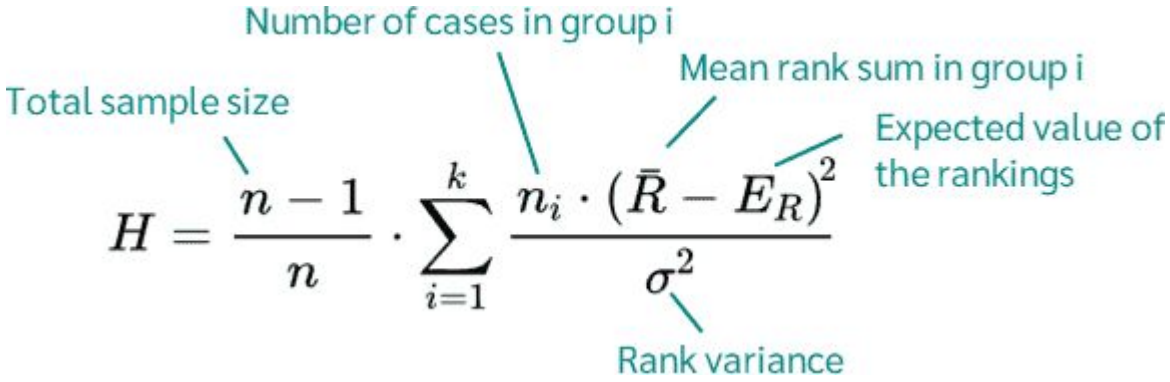
https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_test

Kruskal-Wallis-Test

The null hypothesis (H0) is that the population medians are equal.

The alternative hypothesis (H1) is that the population medians are not equal, or that the population median differs from the population median of one of the other groups.





https://datatab.net/tutorial/kruskal-wallis-test

Kruskal-Wallis-Test

```
head(Data_Krukal,8)
```

A data.frame: 8 × 2

	value	Group
	<dbl>	<fct>
5_15_drySt-28F	94.75000	Fecal.Vdr-/-
20_12_CeSt-28F	59.80000	Cecal.Vdr-/-
1_11_drySt-28F	77.00000	Fecal.Vdr-/-
2_12_drySt-28F	103.27273	Fecal.Vdr-/-
3_13_drySt-28F	85.66667	Fecal.Vdr-/-
4_14_drySt-28F	55.14286	Fecal.Vdr-/-
7_22_drySt-28F	62.75000	Fecal.WT
8_23_drySt-28F	67.66667	Fecal.WT

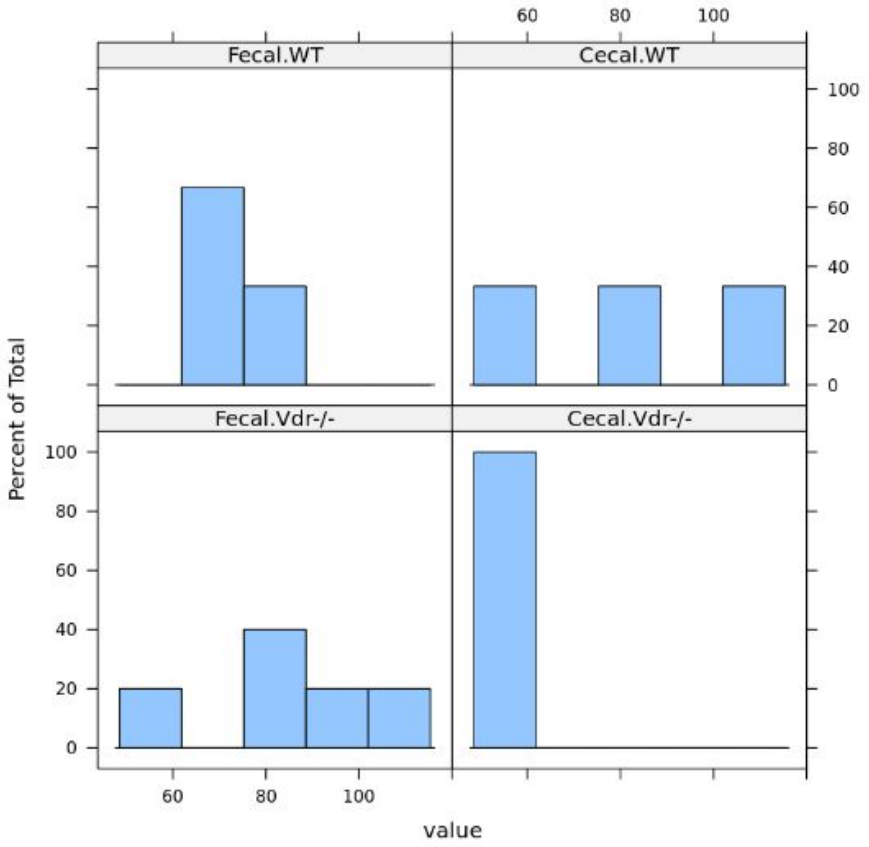
```
tail(Data_Krukal,8)
```

A data.frame: 8 × 2

	value	Group
	<dbl>	<fct>
9_24_drySt-28F	80.50000	Fecal.WT
19_11_CeSt-28F	52.16667	Cecal.Vdr-/-
21_13_CeSt-28F	55.00000	Cecal.Vdr-/-
22_14_CeSt-28F	59.00000	Cecal.Vdr-/-
23_15_CeSt-28F	60.87500	Cecal.Vdr-/-
25_22_CeSt-28F	51.00000	Cecal.WT
26_23_CeSt-28F	112.85714	Cecal.WT
27_24_CeSt-28F	78.05882	Cecal.WT

Kruskal-Wallis-Test

```
histogram(~ value|Group, data=Data_Krukal, layout=c(2,2))
```



H0: the medians of each group are the same
HA: at least one of the groups has a different median

```
# kruskal wallis test of richness
kruskal.test(value ~ Group, data = Data_Krukal)
```

```
        Kruskal-Wallis rank sum test

data:  value by Group
Kruskal-Wallis chi-squared = 5.2353, df = 3, p-value = 0.1554
```

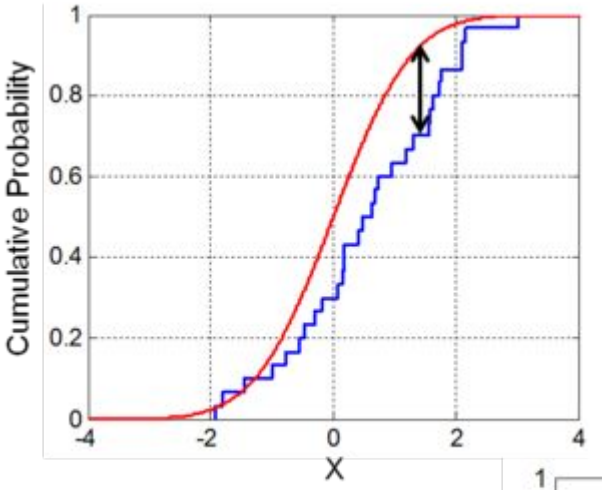
Kolmogorov-Smirnov Test (KS test or K-S test)

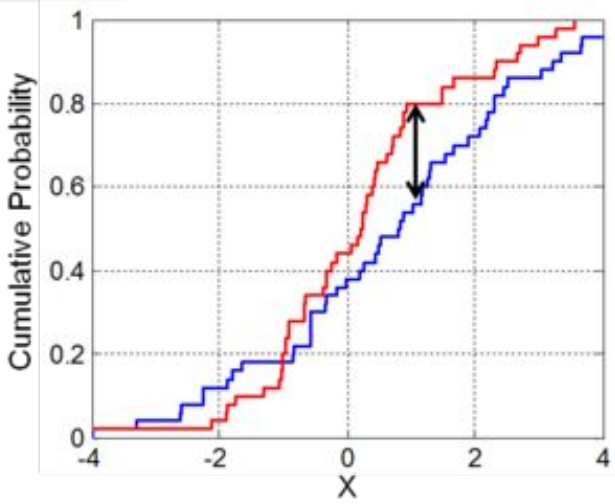
The Kolmogorov Smirnov test (KS test or K-S test) is used to

- compare a **sample distribution** with a reference probability distribution
- compare **two sample distributions**

Null Hypothesis : The sample follows a specified distribution.

Alternative Hypothesis: The sample does not follow the specified distribution.

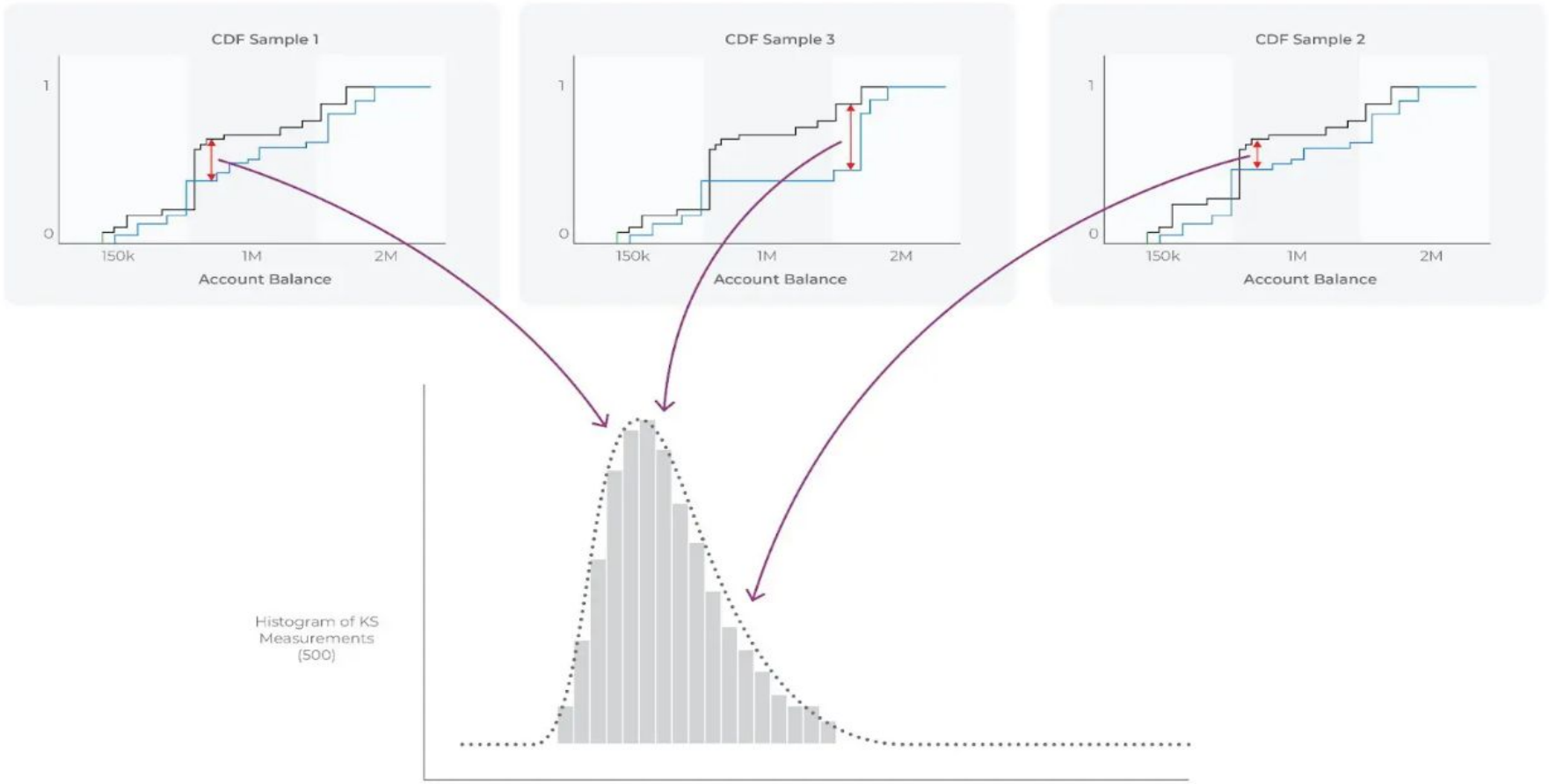




$$D_n = \sup_x \mid \mathrm{F}_n(x) - \mathrm{F}(x) \mid$$

https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Kolmogorov-Smirnov Test (KS test or K-S test)



500 samples of K-S test statistic

$$D_n = \sup_x | F_n(x) - F(x) |$$

https://arize.com/blog-course/kolmogorov-smirnov-test/

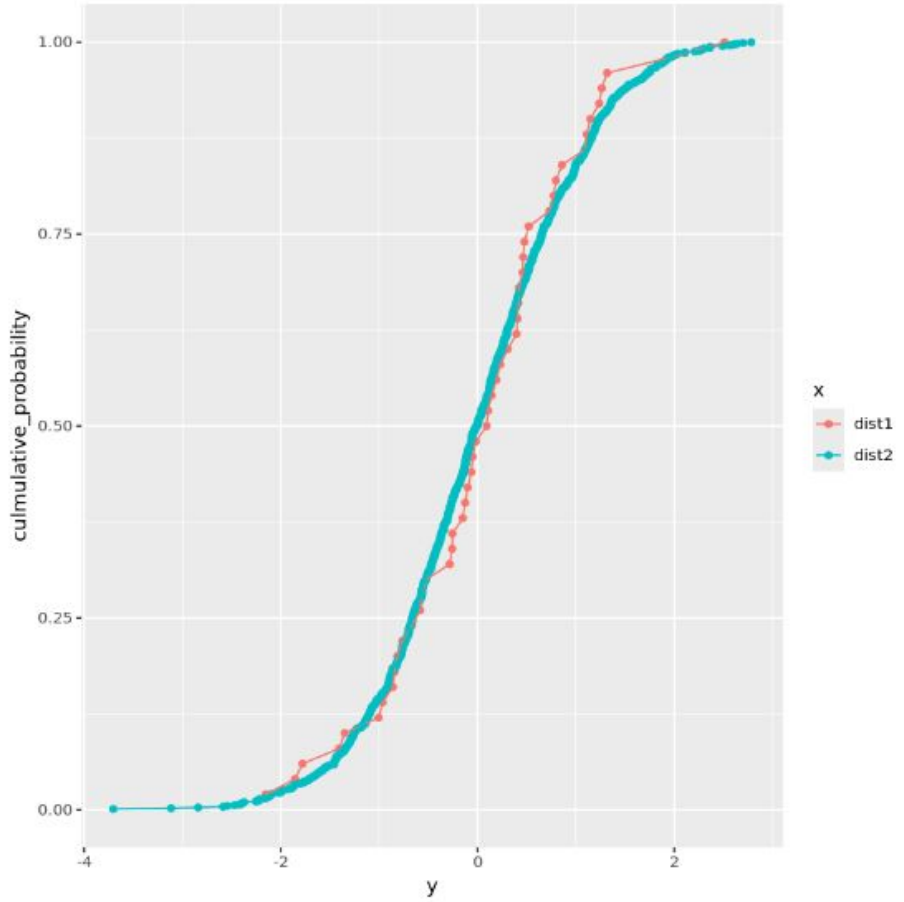
Kolmogorov-Smirnov Test (KS test or K-S test)

```r
# Make this example reproducible
set.seed(0)
```

```r
# Generate dataset of 100 values
# that follow a Standard Normal distribution
data <- rnorm(n=200, mean=0, sd=1)
```

```r
# Perform Kolmogorov-Smirnov test
ks.test(data, 'pnorm')
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data
D = 0.035916, p-value = 0.9587
alternative hypothesis: two-sided
```



https://www.statology.org/kolmogorov-smirnov-test-r/

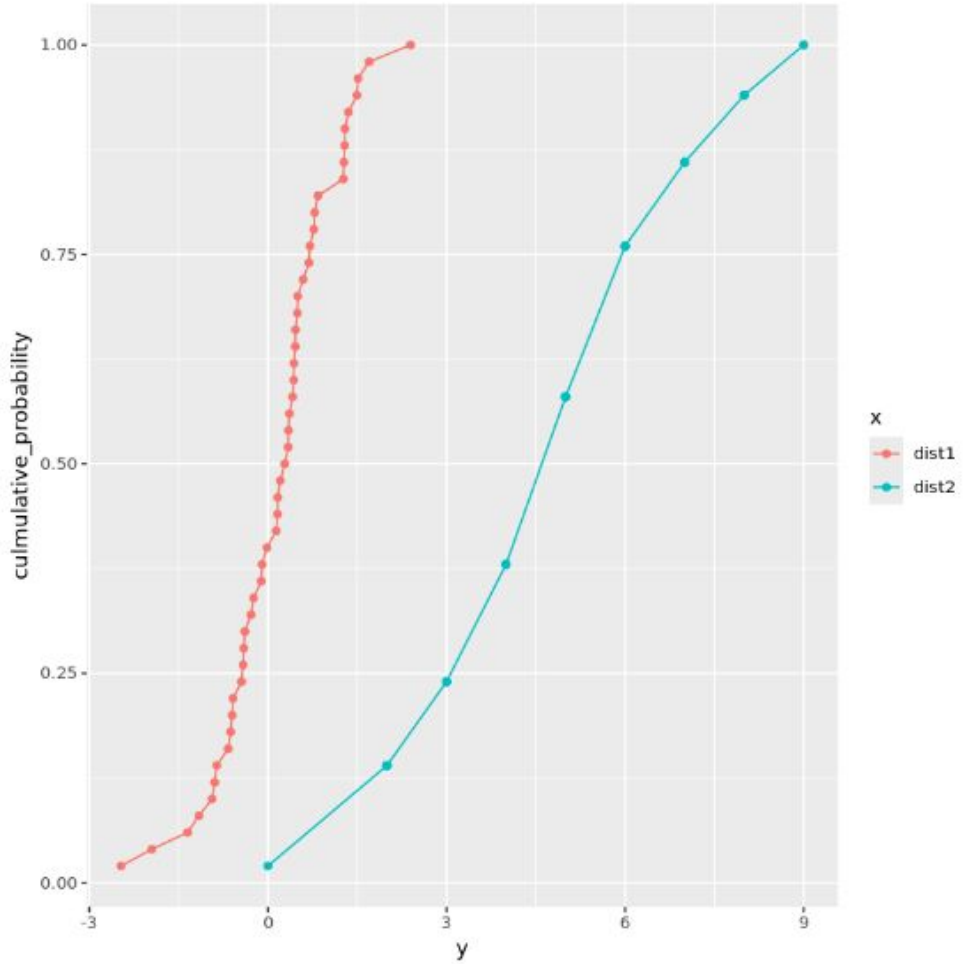
Kolmogorov-Smirnov Test (KS test or K-S test)

```r
# Make this example reproducible
set.seed(0)
```

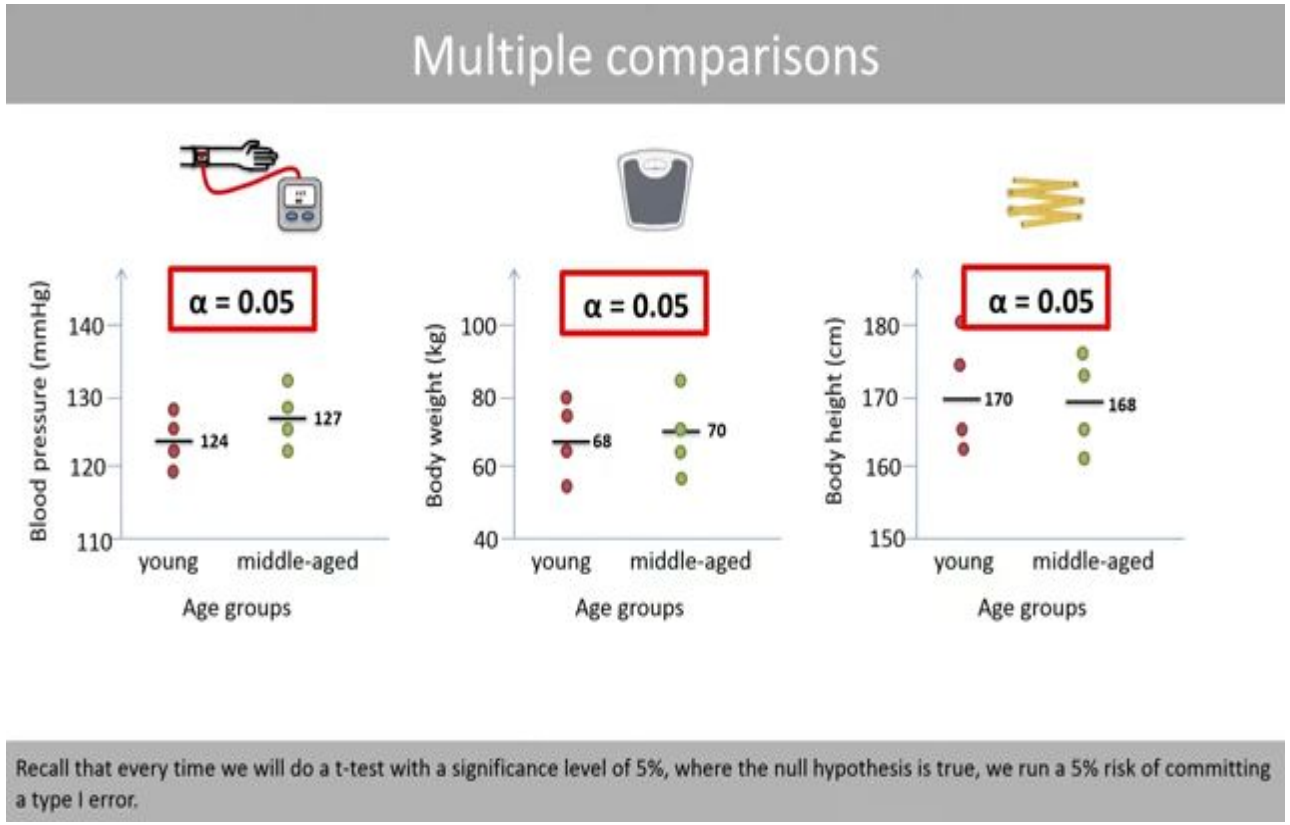
```r
# Generate dataset of 100 values
# that follow a Standard Normal distribution
data1 <- rnorm(n=50, mean=0, sd=1)
data2 <- rpois(n=50, lambda=5)
```

```r
# Perform Kolmogorov-Smirnov test
ks.test(data1, data2)
```

```
        Exact two-sample Kolmogorov-Smirnov test

data:  data1 and data2
D = 0.94, p-value < 2.2e-16
alternative hypothesis: two-sided
```



False Discovery Rate



Family-Wise Error Rate

$$1 - (1 - \alpha)^{\text{Number of tests}}$$

1 Test $1 - (1 - 0.05)^1 = 1 - (0.95)^1 = 0.05$ or 5%

2 Test ~ $1 - (1 - 0.05)^2 = 1 - (0.95)^2 = 0.10$ or 10%

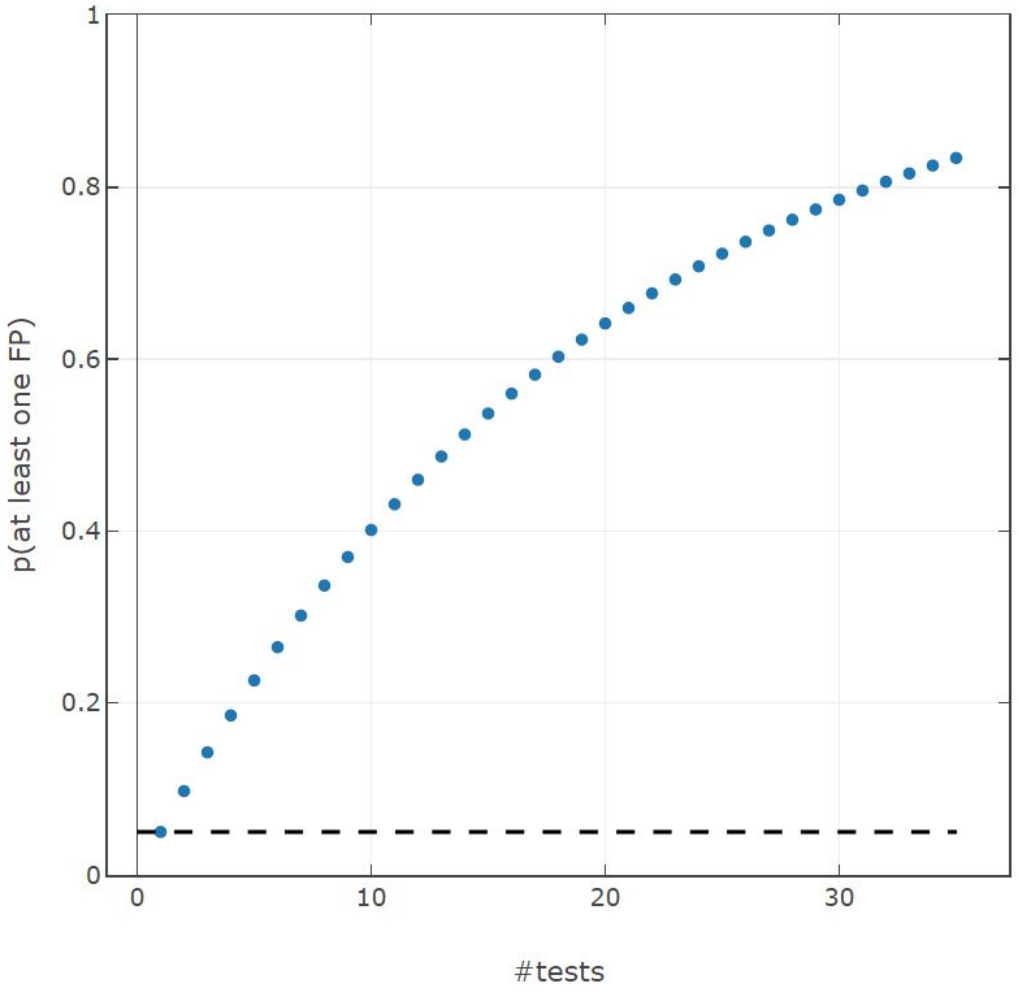
3 Test ~ $1 - (1 - 0.05)^3 = 1 - (0.95)^3 = 0.14$ or 14%

4 Test ~ $1 - (1 - 0.05)^4 = 1 - (0.95)^4 = 0.19$ or 19%

5 Test ~ $1 - (1 - 0.05)^5 = 1 - (0.95)^5 = 0.23$ or 23%

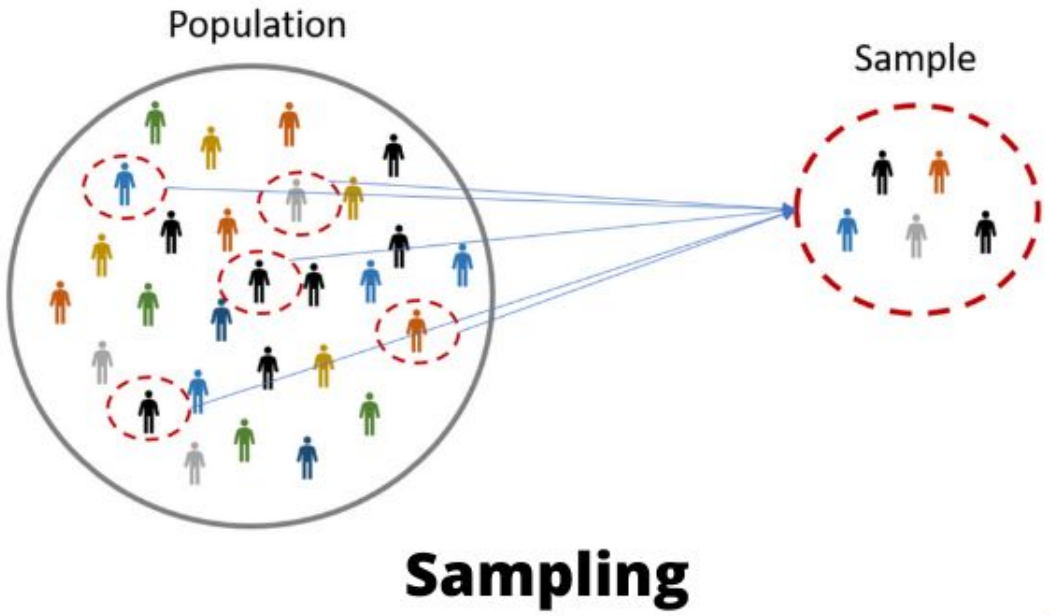
https://www.youtube.com/watch?v=8CuFKY45UXQ
https://medium.com/illumination/decoding-data-success-a-comprehensive-guide-to-a-b-testing-analysis-part-3-b13fc7314e07

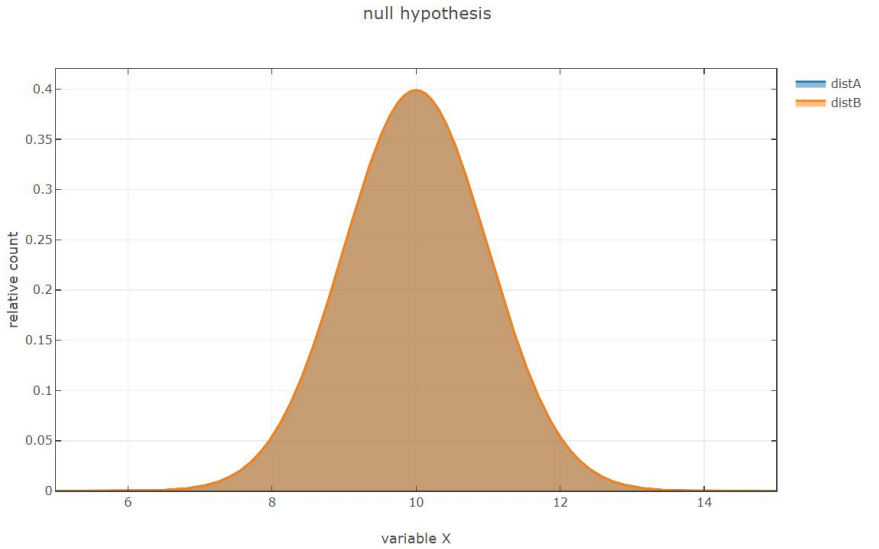
False Discovery Rate

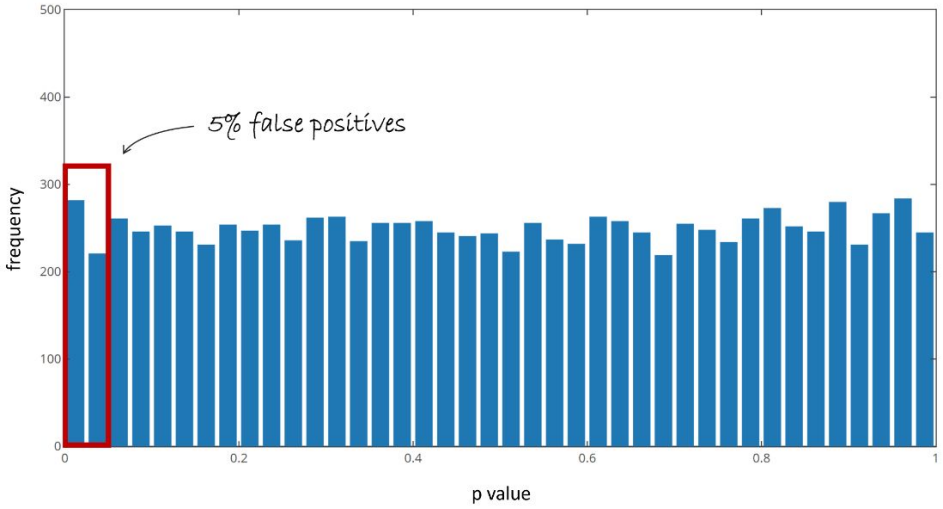


https://fslab.org/blog/posts/q-value.html

False Discovery Rate

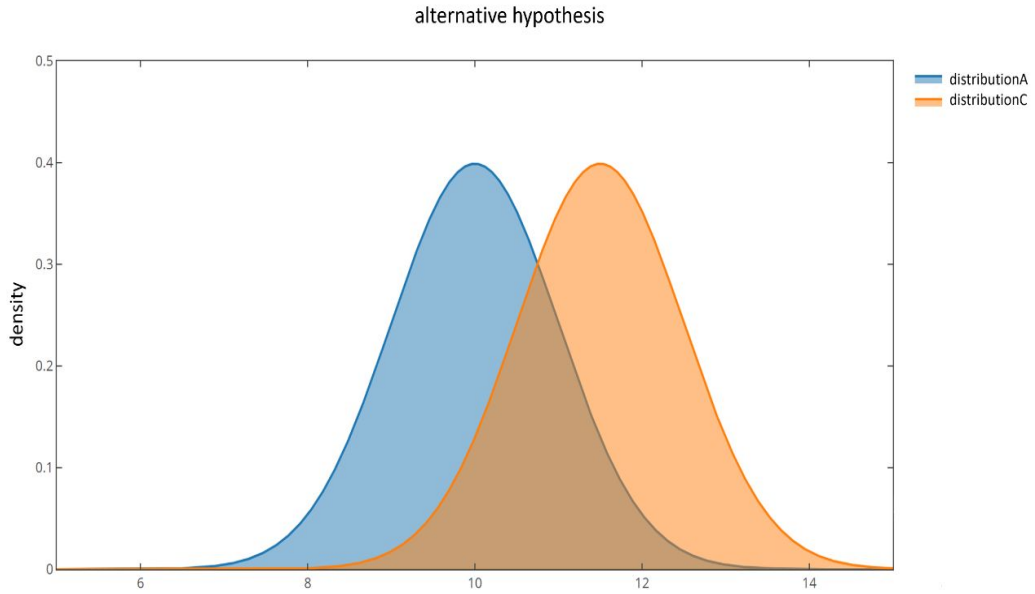


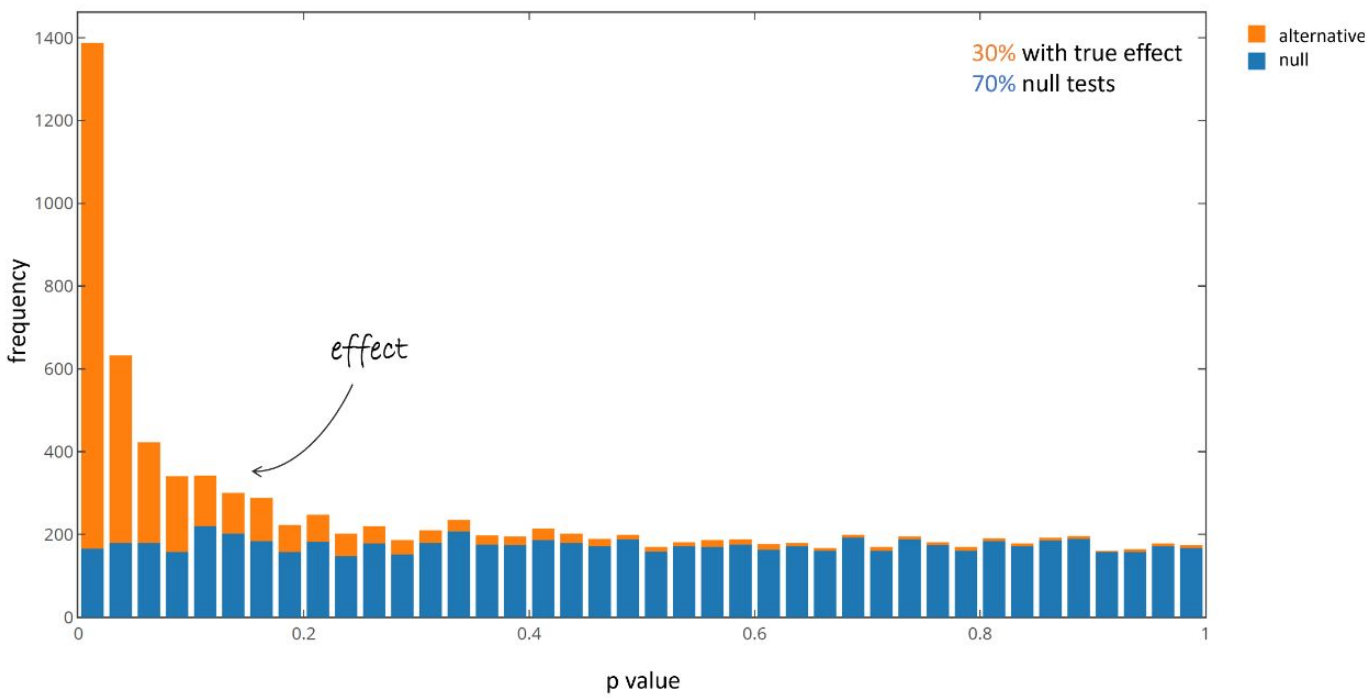




https://fslab.org/blog/posts/q-value.html
https://hotcubator.com.au/research/a-complete-guide-to-sampling-techniques/

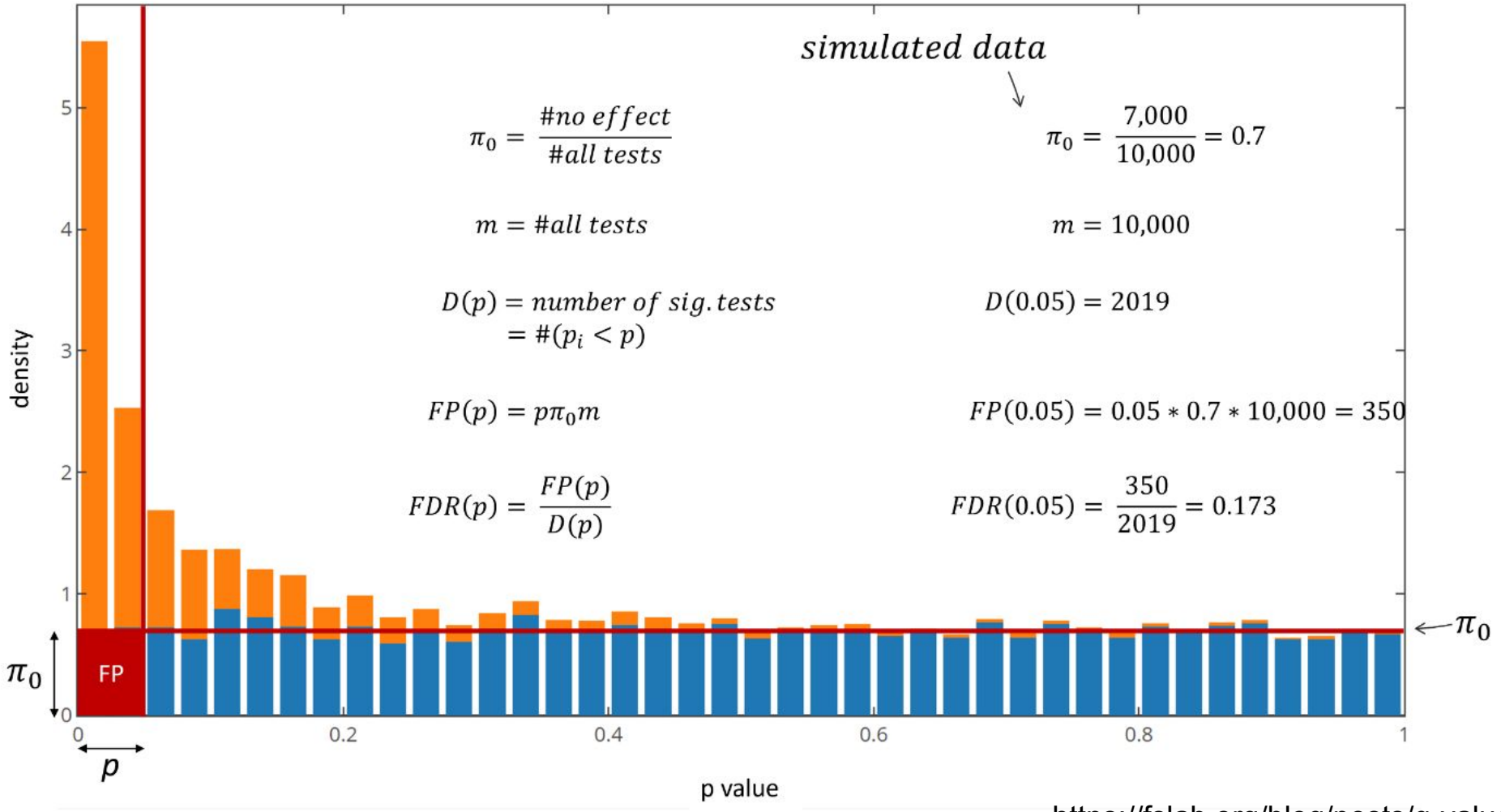
False Discovery Rate





https://fslab.org/blog/posts/q-value.html

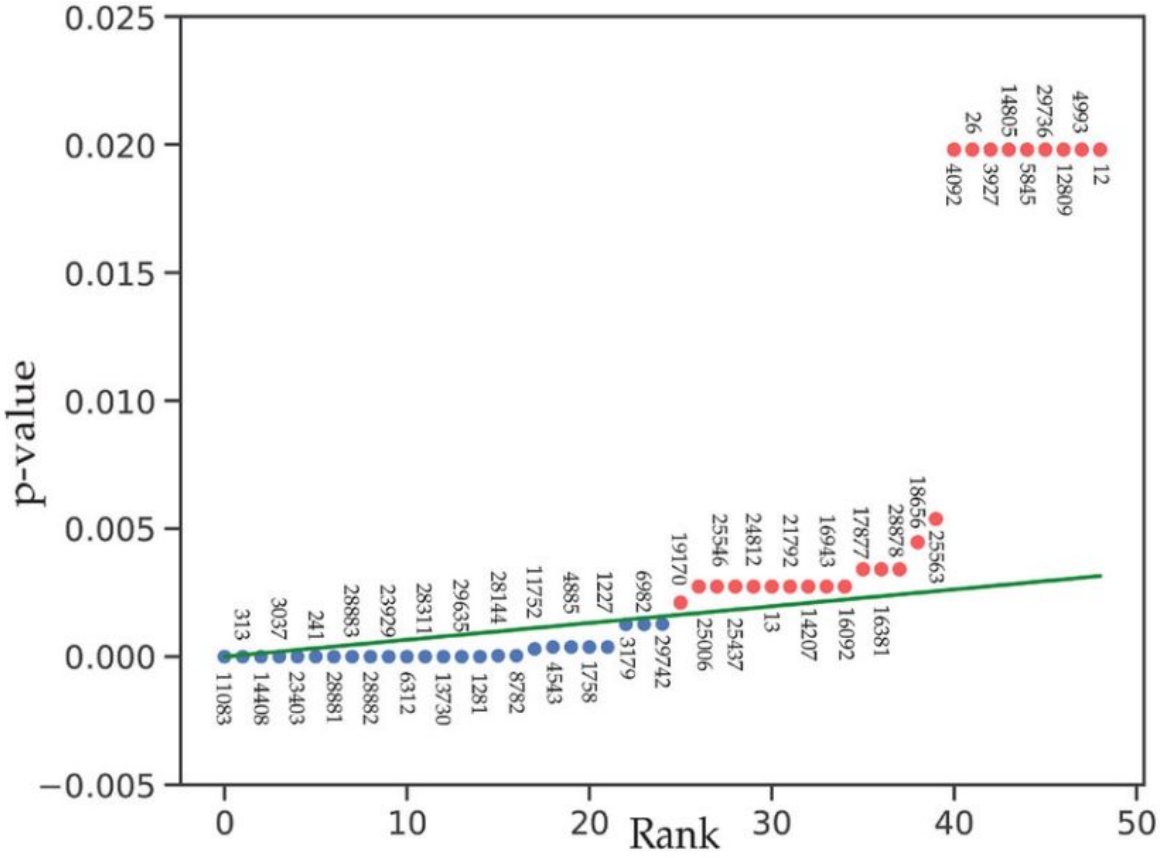
False Discovery Rate



https://fslab.org/blog/posts/q-value.html

https://physiology.med.cornell.edu/people/banfelder/qbio/resources_2008/1.5_Bonferroni_FDR.pdf

False Discovery Rate



Benjamini Hochberg Method

DOI:10.1021/acs.jpclett.0c02765

False Discovery Rate

```
pvalues<-c(0.01,0.001, 0.05, 0.20, 0.15, 0.15)

fdrs<-p.adjust(pvalues, method="BH") # Benjamini-Hochberg
print(fdrs)
```

```
[1] 0.030 0.006 0.100 0.200 0.180 0.180
```

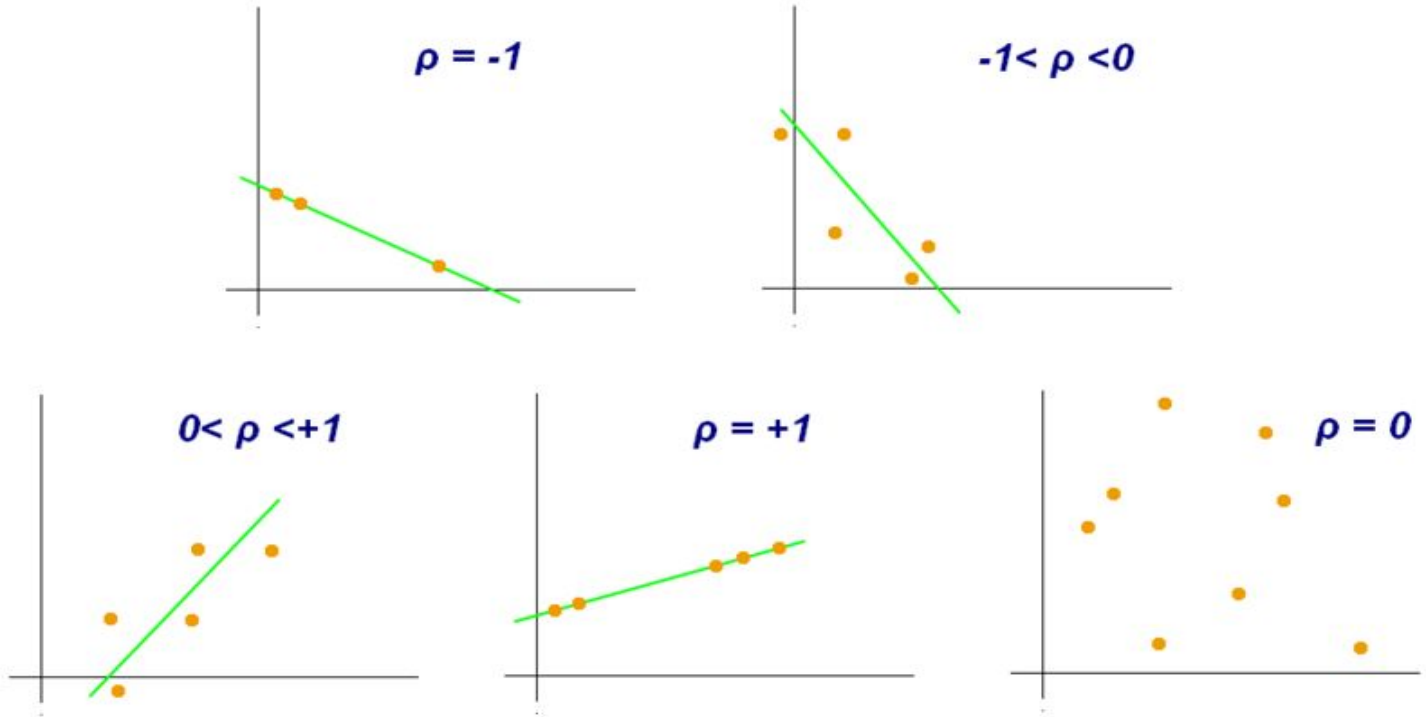
https://www.r-bloggers.com/2023/07/the-benjamini-hochberg-procedure-fdr-and-p-value-adjusted-explained/

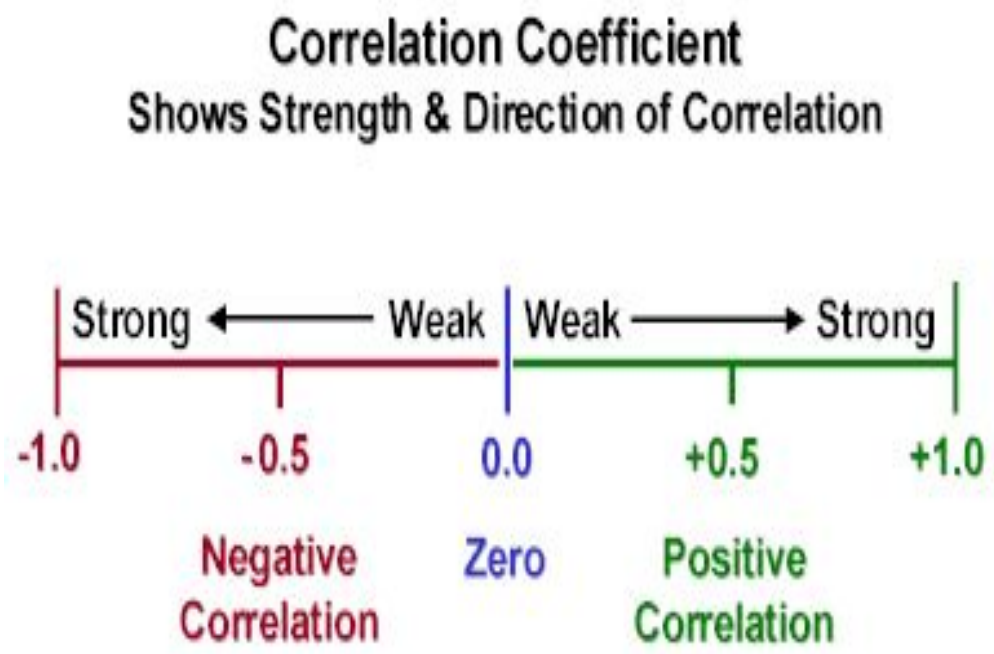
4. Correlation and linear regression

Correlation

- Determine the **association** between the two quantitative variables.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

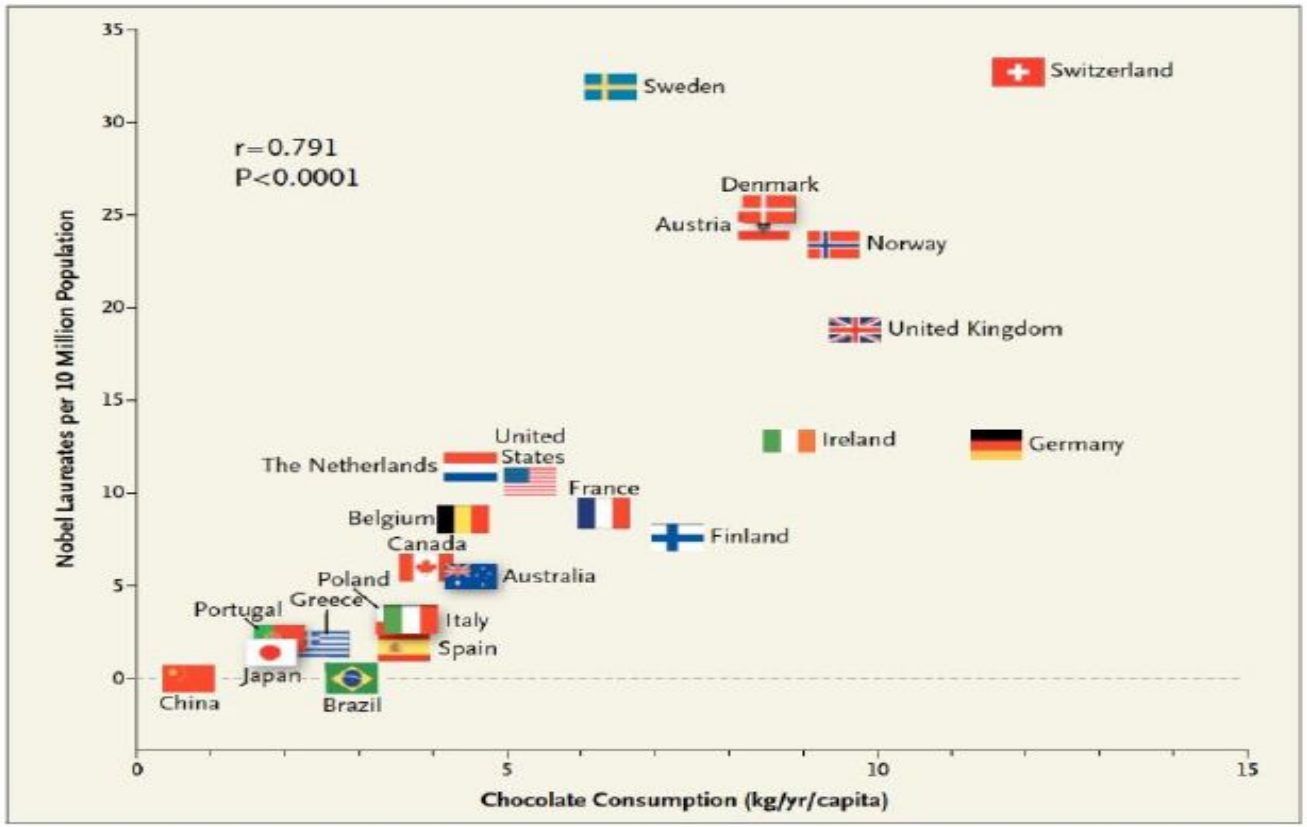




4. Correlation and linear regression

Correlation

Correlation indicates association, not causation



Correlation and linear regression

Linear regression

- A supervised machine learning technique
- Establish a relationship between a dependent variable and one or more independent variables.
- Determine if there is a linear relationship between two or more variables and how strong that relationship is.

Correlation and linear regression

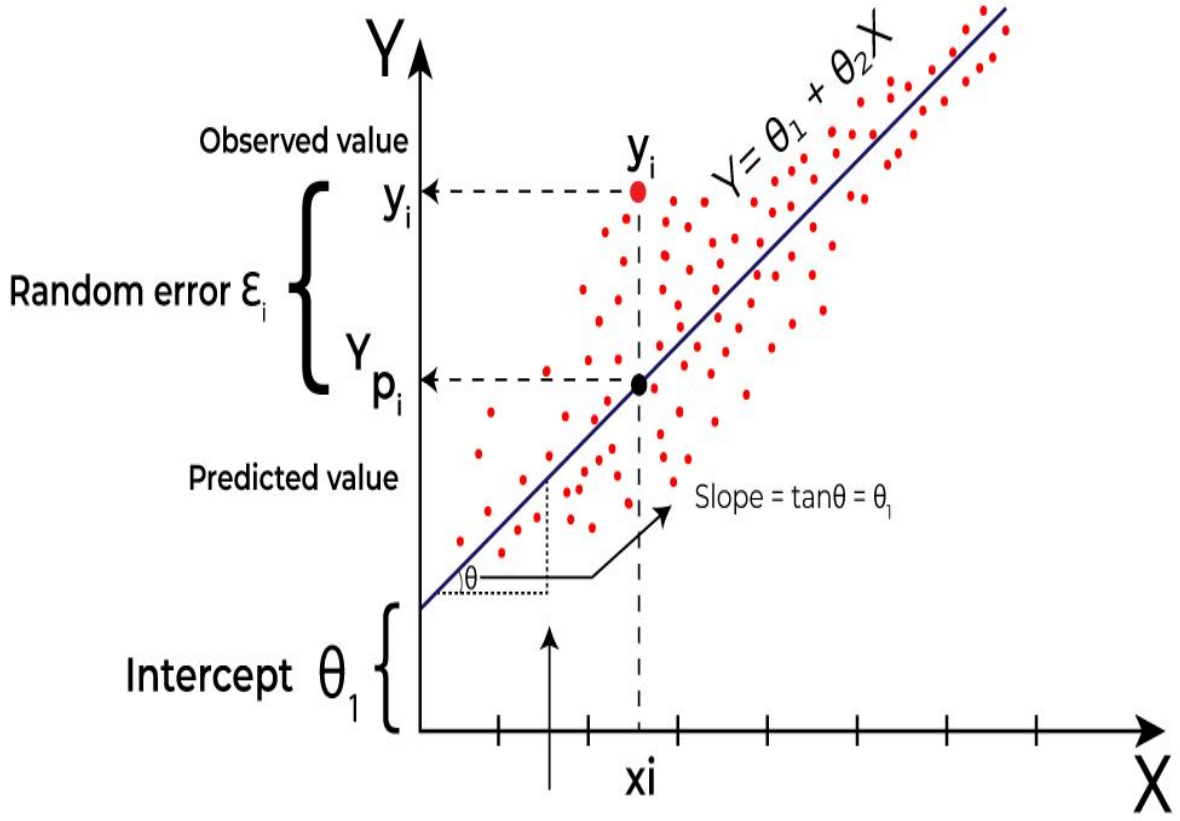
Linear regression

Linear Regression Equation

$$Y = a + bx$$

$$a = \frac{[(\Sigma y)(\Sigma x^2) - (\Sigma y)(\Sigma xy)]}{[n(\Sigma x^2) - (\Sigma x)^2]}$$

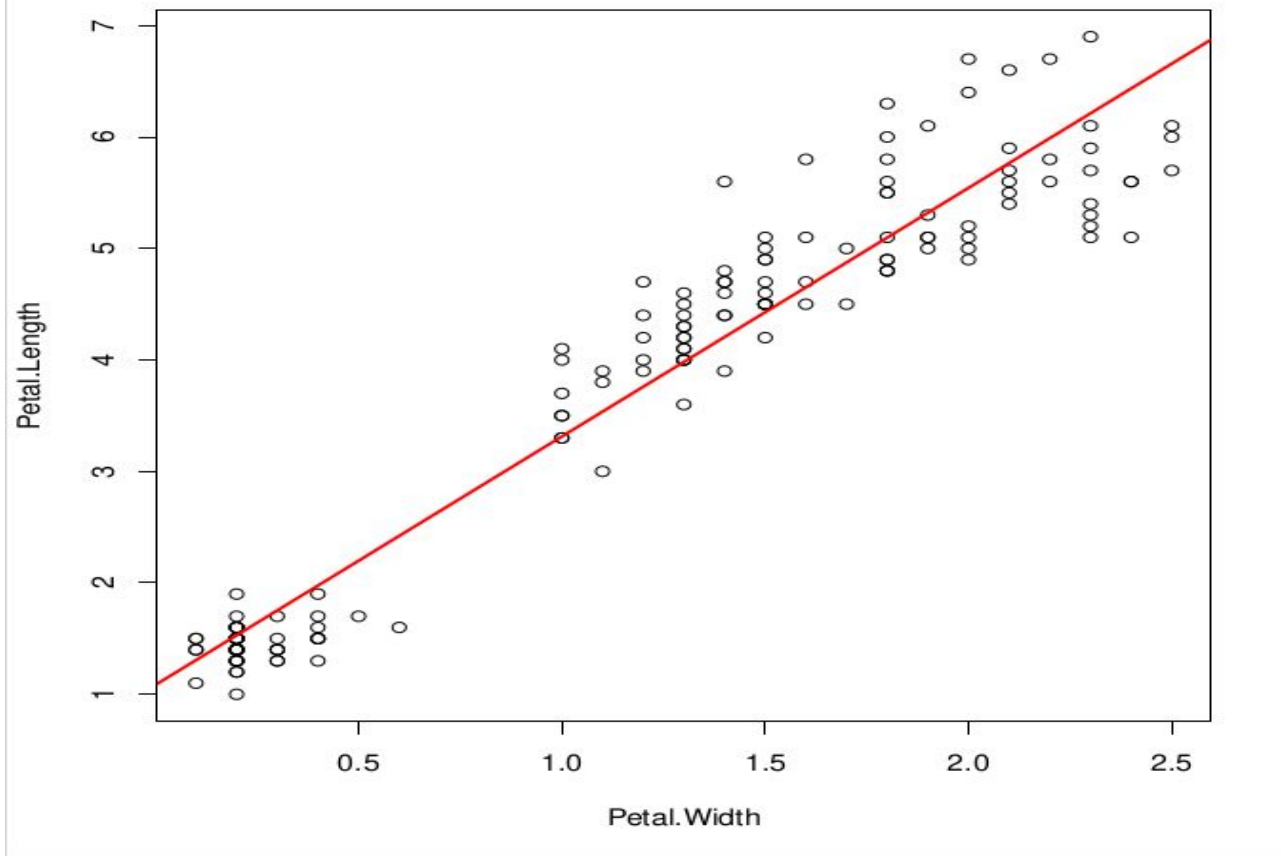
$$b = \frac{[n(\Sigma xy) - (\Sigma x)(\Sigma y)]}{[n(\Sigma x^2) - (\Sigma x)^2]}$$



Correlation and linear regression

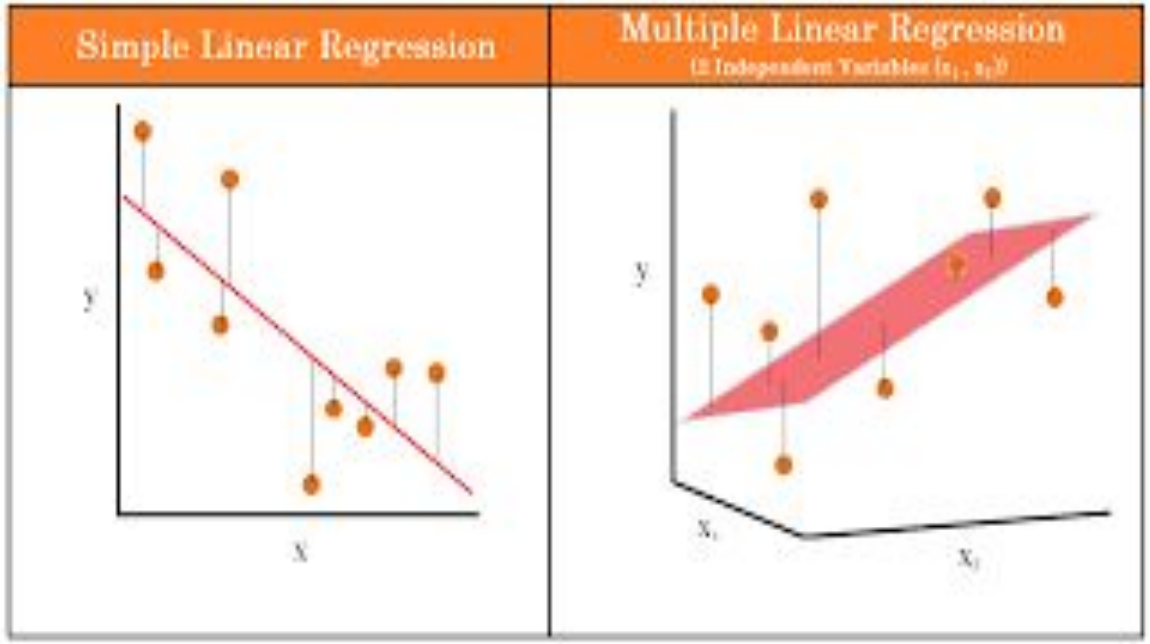
Linear regression

```
Call:
lm(formula = Petal.Length ~ Petal.Width, data = dt.iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33542 -0.30347 -0.02955  0.25776  1.39453

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.08356    0.07297   14.85   <2e-16 ***
Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4782 on 148 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```



Correlation and linear regression

Multiple linear regression



One Predictor Model

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Nonrandom or Systematic Component

Random Component

Multiple Predictor Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_q x_q + \varepsilon$$

Where

Y is the outcome value

β_0 is the intercept

ε is the error aka residual

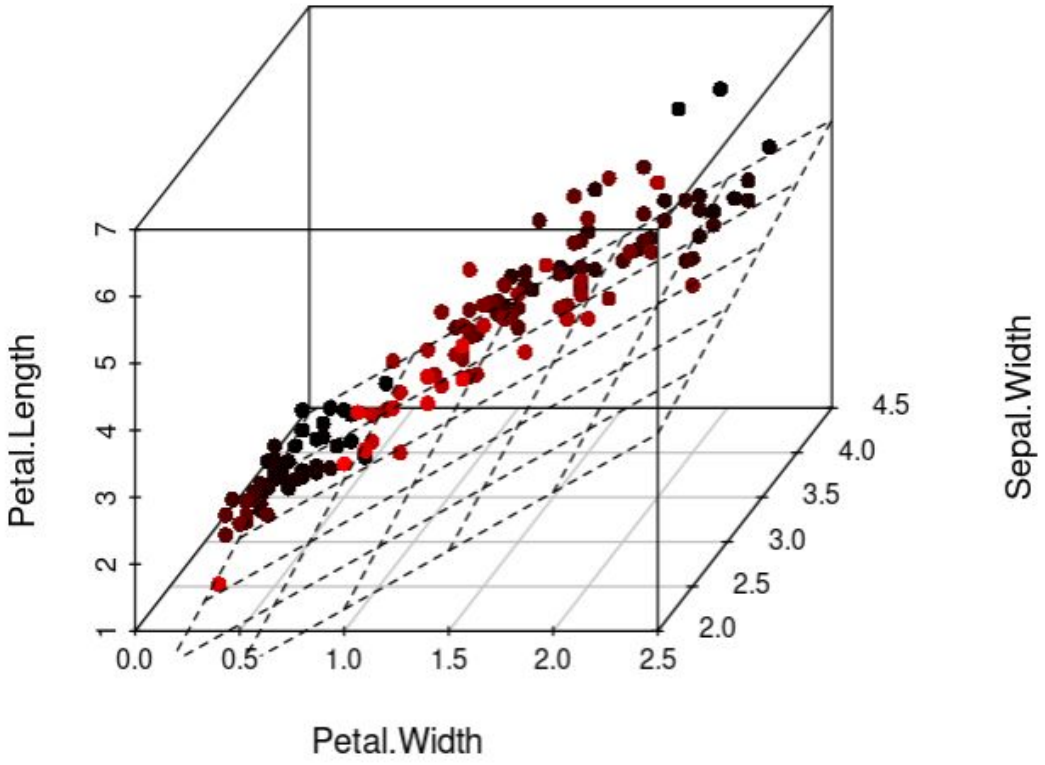
$x_{1..q}$ is the value of predictor variable

$\beta_{1..q}$ is the slope coefficient

Correlation and linear regression

Multiple linear regression

```
>   summary(model2)

Call:
lm(formula = Petal.Length ~ Petal.Width + Sepal.Width, data = dt.iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33753 -0.29251 -0.00989  0.21447  1.24707

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.25816    0.31352   7.203 2.84e-11 ***
Petal.Width  2.15561    0.05283  40.804  < 2e-16 ***
Sepal.Width -0.35503    0.09239  -3.843  0.00018 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4574 on 147 degrees of freedom
Multiple R-squared:  0.9338,    Adjusted R-squared:  0.9329
F-statistic:  1036 on 2 and 147 DF,  p-value: < 2.2e-16
```



Summary

Summary

Type		Name	Function in R
Descriptive statistics		Variance	var()
		Covariance	cov()
		Correlation	cor()
Inferential statistics	Parametric tests	T-test	t.test()
		ANOVA	aov()
	Non-parametric tests	Wilcoxon test	wilcox.test()
		Kruskal-Wallis test	kruskal.test()
		Chi-squared test	chisq.test()
	Normality tests	Shapiro-Wilk Test	shapiro.test()
		Kolmogorov-Smirnov Test	ks.test()
Supervised machine learning		Linear Regression	lm()