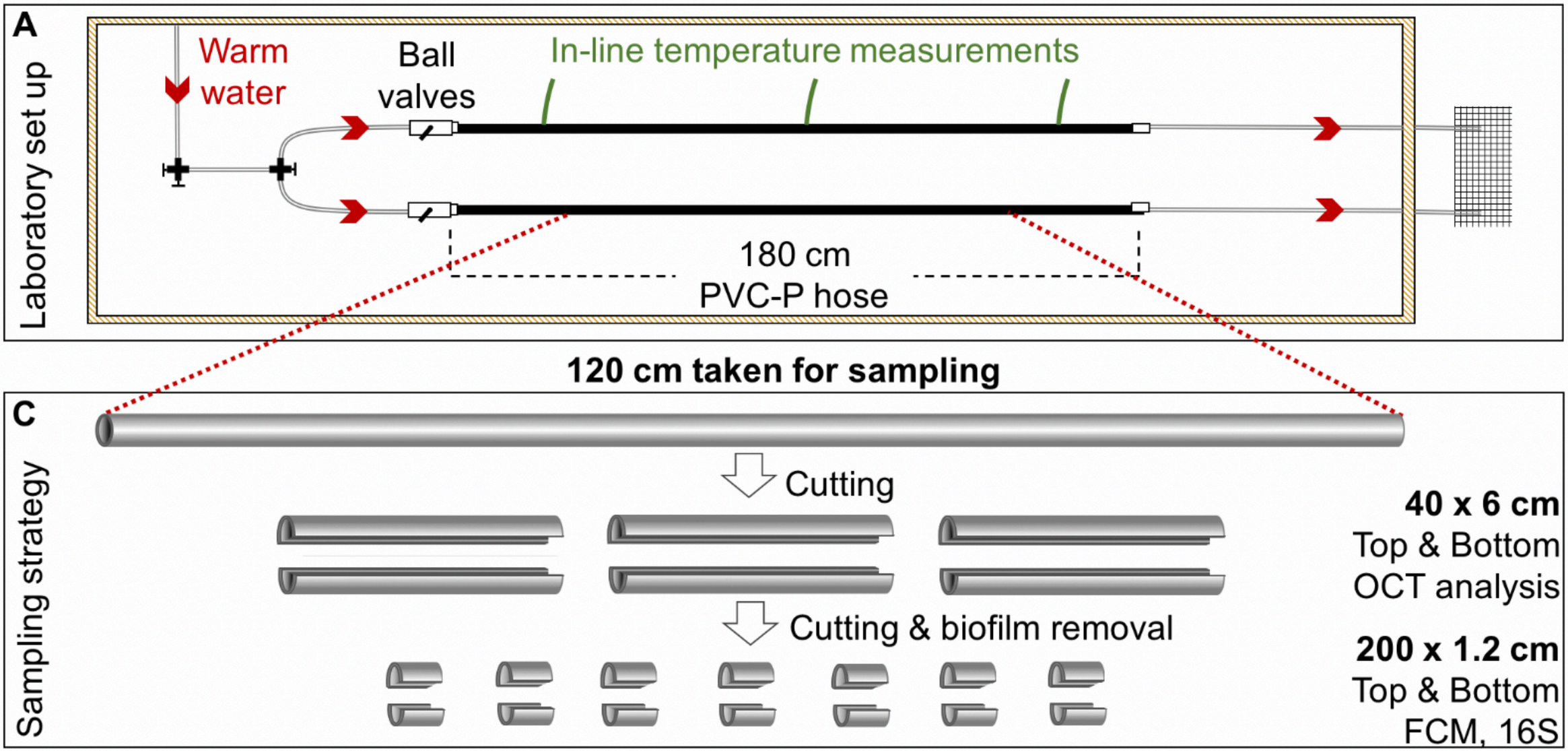
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

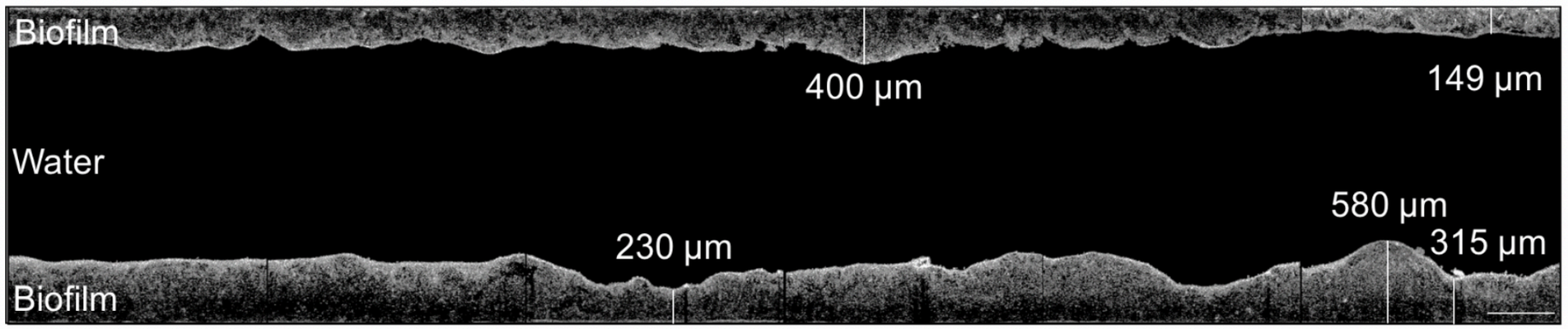
Genetic Diversity Centre Zurich

Data Visualisation and Basic Statistics

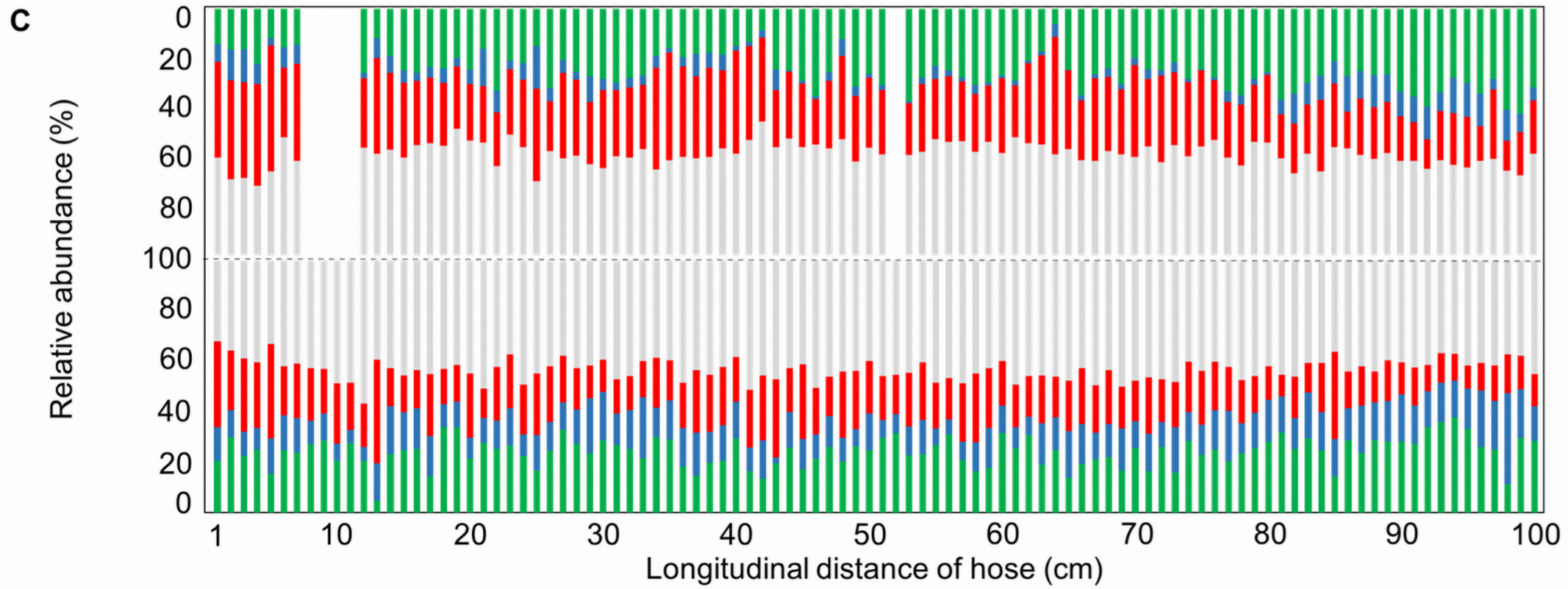
Jean-Claude Walser
jean-claude.walser@env.ethz.ch

Anne C. Roulin
anne.roulin@botinst.uzh.ch

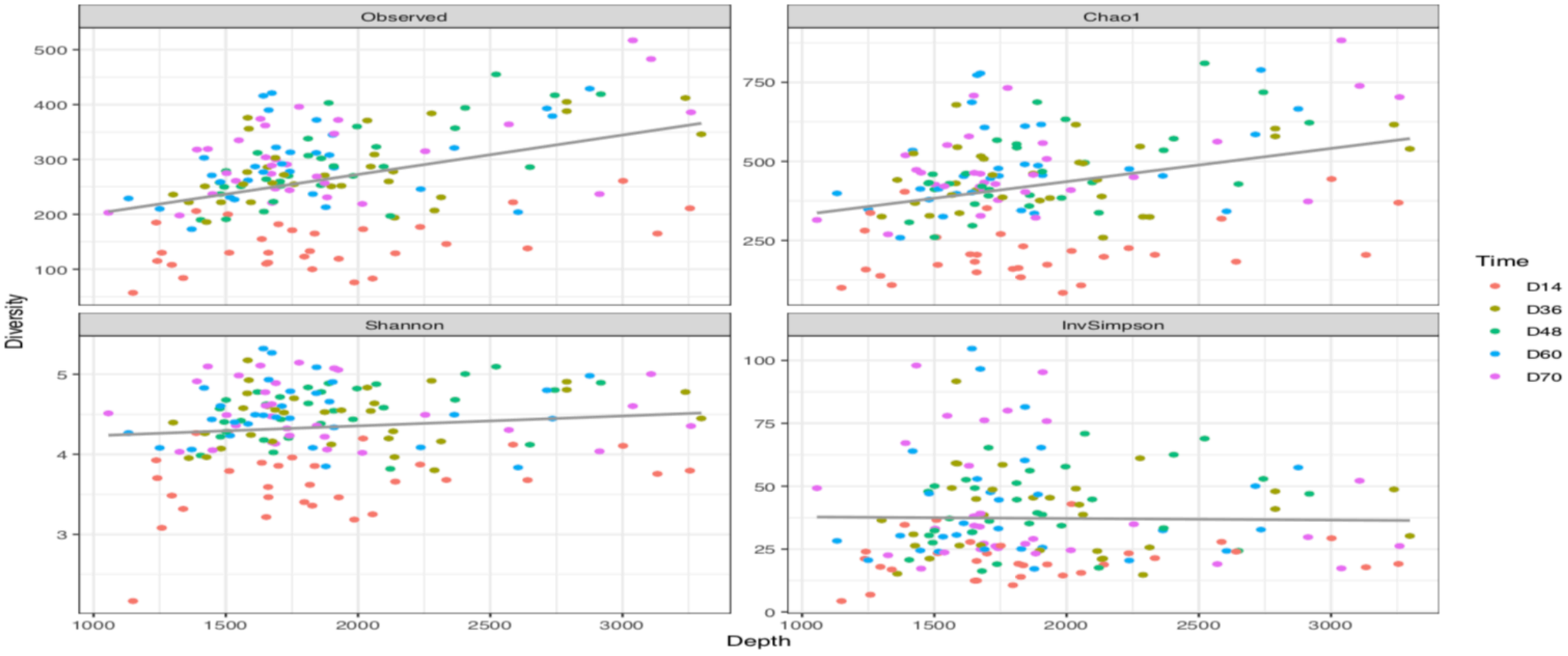

A
Laboratory set up
Warm water
Ball valves
In-line temperature measurements
180 cm
PVC-P hose
120 cm taken for sampling
C
Sampling strategy
Cutting
Cutting & biofilm removal
40 x 6 cm
Top & Bottom
OCT analysis
200 x 1.2 cm
Top & Bottom
FCM, 16S

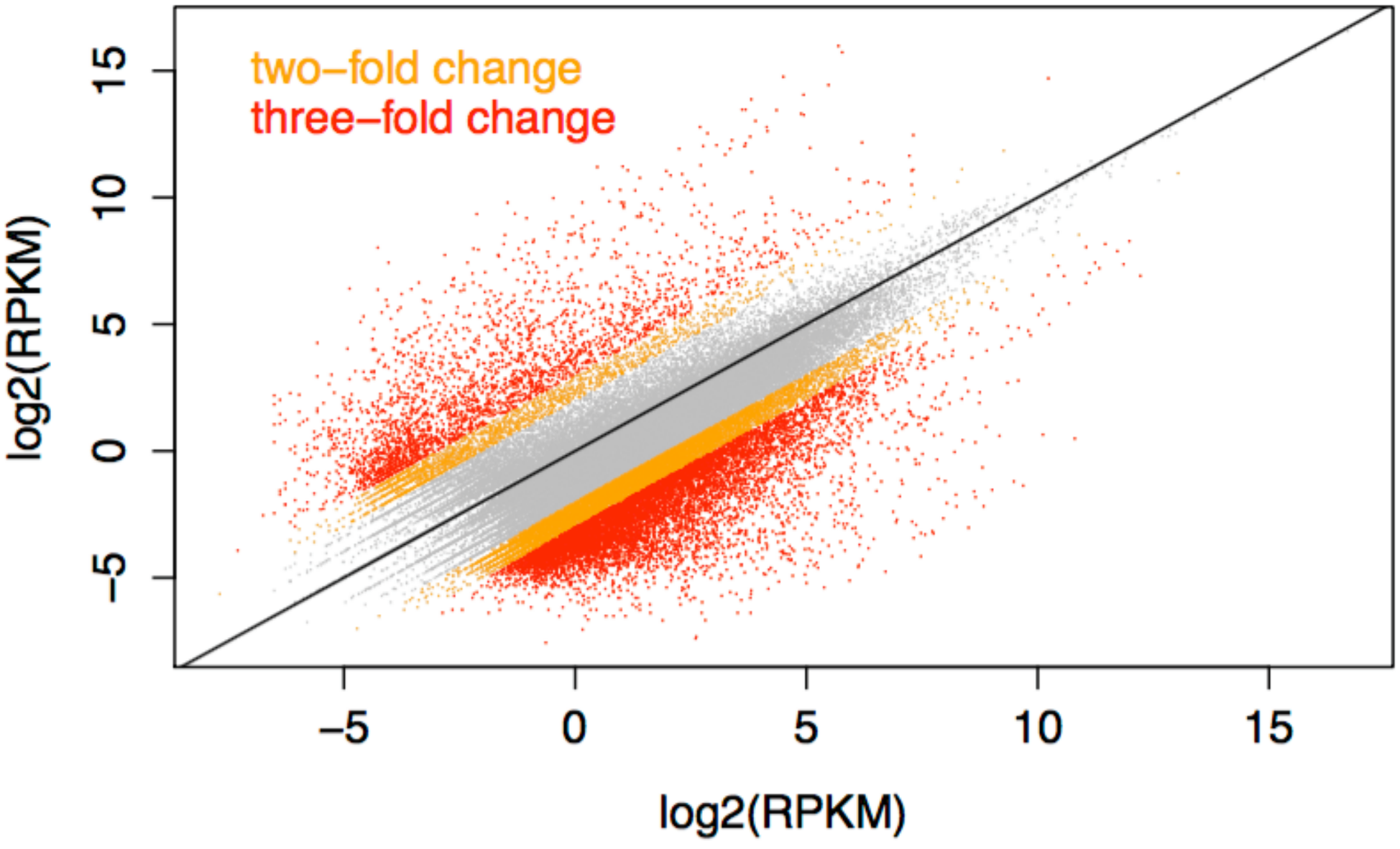


The averaged thickness of 1.2 cm-sections ranged between 150 – 750 µm with an overall average of 319 ± 111 µm (n = 200). On large-scale, the biofilm was significantly thicker at the bottom (386 ± 117 µm, n = 100) compared to the top (252 ± 44 µm, n = 100) (**t-test, p < 0.05**).



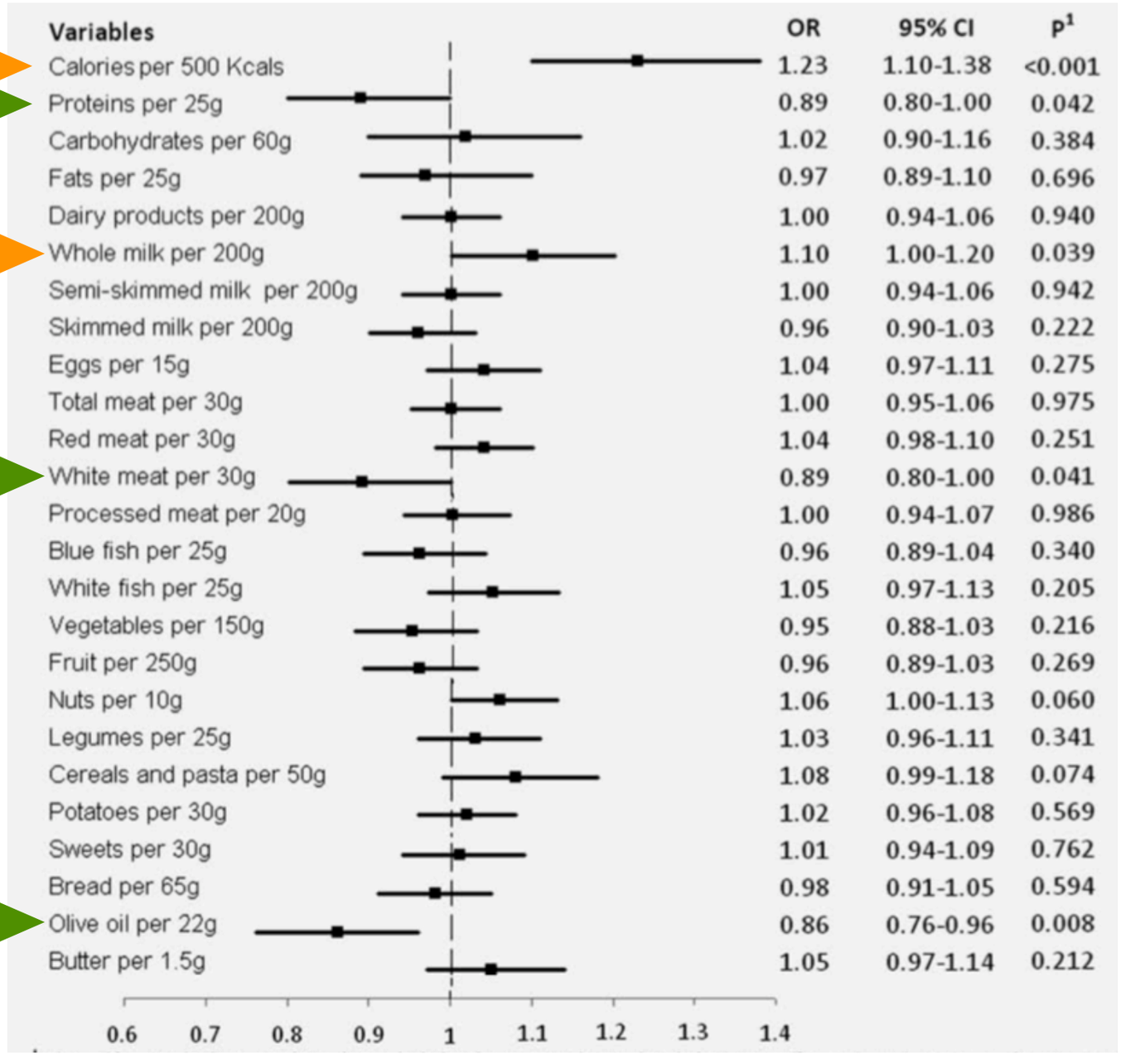
Interestingly, only few dominant taxa made up for the majority of the community composition. In the top part of the hose, 12 taxa accounted for 92 % of the total community. A similar situation was identified for the dominant taxa in the bottom biofilm, with eleven taxa covering 91 %. Dominant taxa were similar between top and bottom, with only three out of 23 being different.

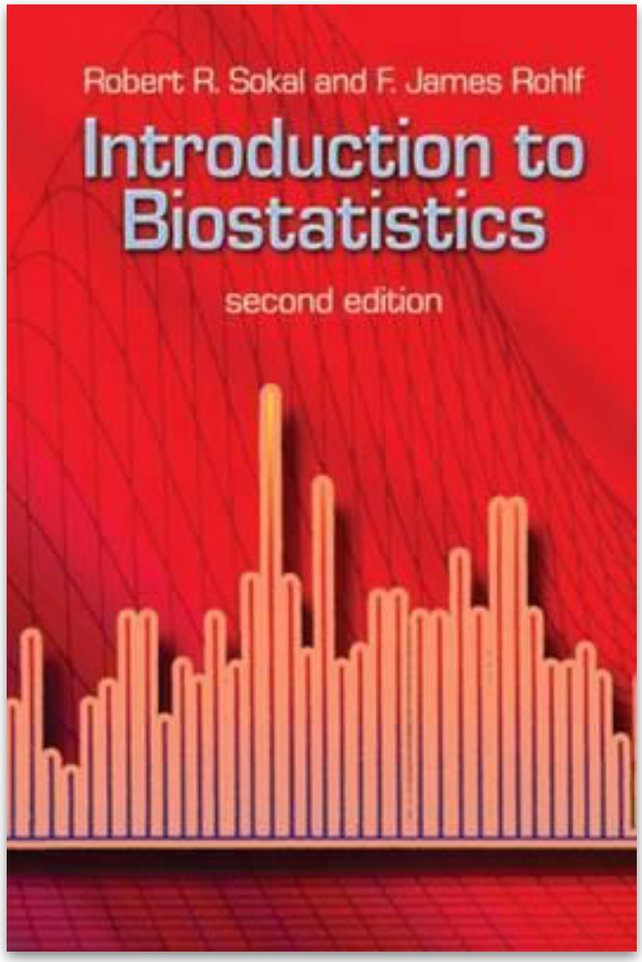

Observed
Chao1
Shannon
InvSimpson
Diversity
Depth
Time
D14
D36
D48
D60
D70

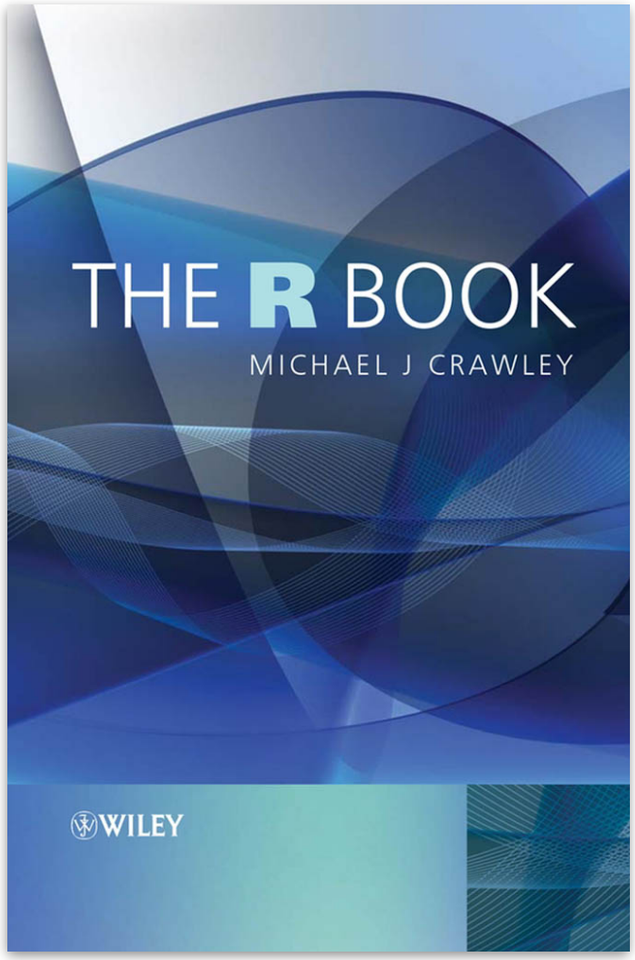

Female: Brain/Antenna
two–fold change
three–fold change
log2(RPKM)
15
10
5
0
−5
−5
0
5
10
15
log2(RPKM)

Calorie intake, olive oil consumption and mammographic density among Spanish women

Nicolás García-Arenzana[1,2], Eva María Navarrete-Muñoz[3,4], Virginia Lope[1,4], Pilar Moreo[5], Carmen Vidal[6], Soledad Laso-Pablos[7,8], Nieves Ascunce[4,9], Francisco Casanova-Gómez[10], Carmen Sánchez-Contador[11], Carmen Santamariña[12], Nuria Aragonés[1,4], Beatriz Pérez Gómez[1,4], Jesús Vioque[3,4] and Marina Pollán[1,4]

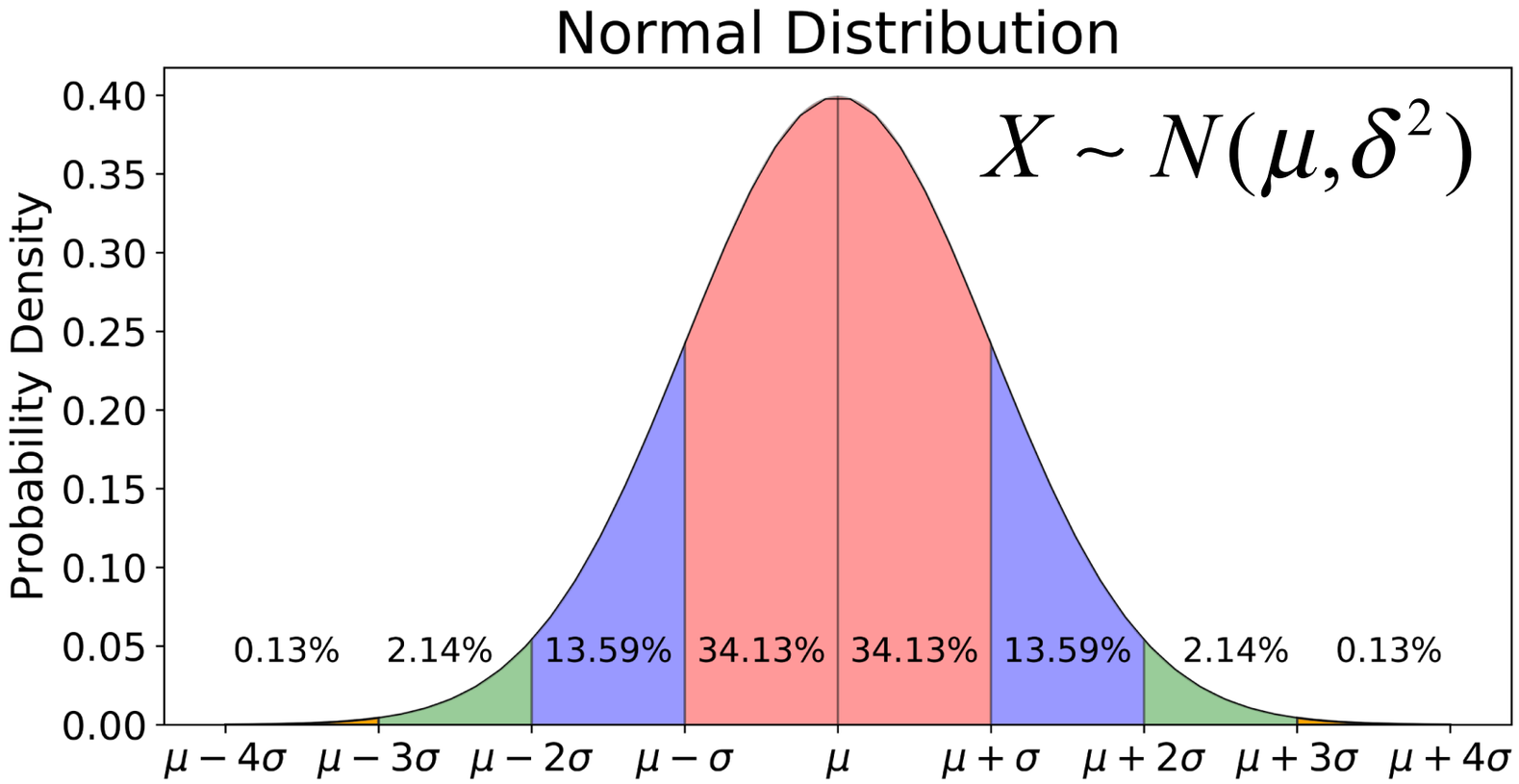



Robert R. Sokal and F. James Rohlf
Introduction to Biostatistics
second edition


THE R BOOK
MICHAEL J CRAWLEY
WILEY

Probability Distributions

Binomial / Normal / Poisson



μ: mean (also median and mode)

δ: standard deviation

δ^2: variance

The **normal distribution** is one of the most important and most widely used distribution in statistics. It is sometimes called the "bell curve".

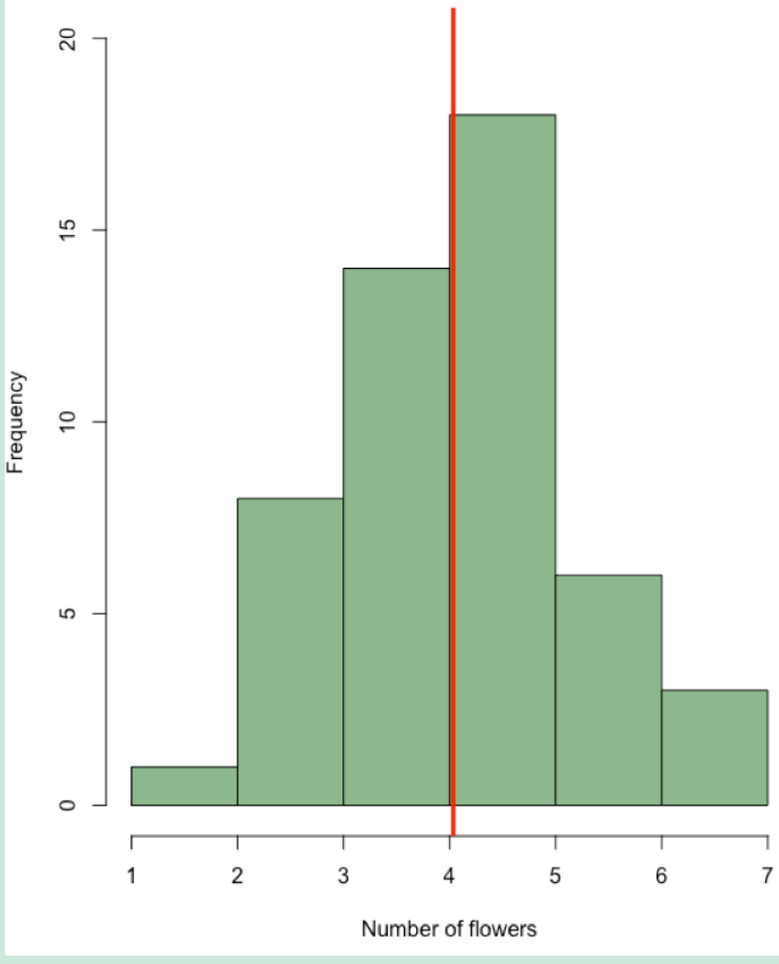
Why does it matter?

Parametric statistics is a branch of statistics which assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters. **Most well-known elementary statistical methods (e.g. t-test, anova) are parametric.**

Checking the distribution of your data and the assumptions of the statistical test is the **first step** of the analysis!!

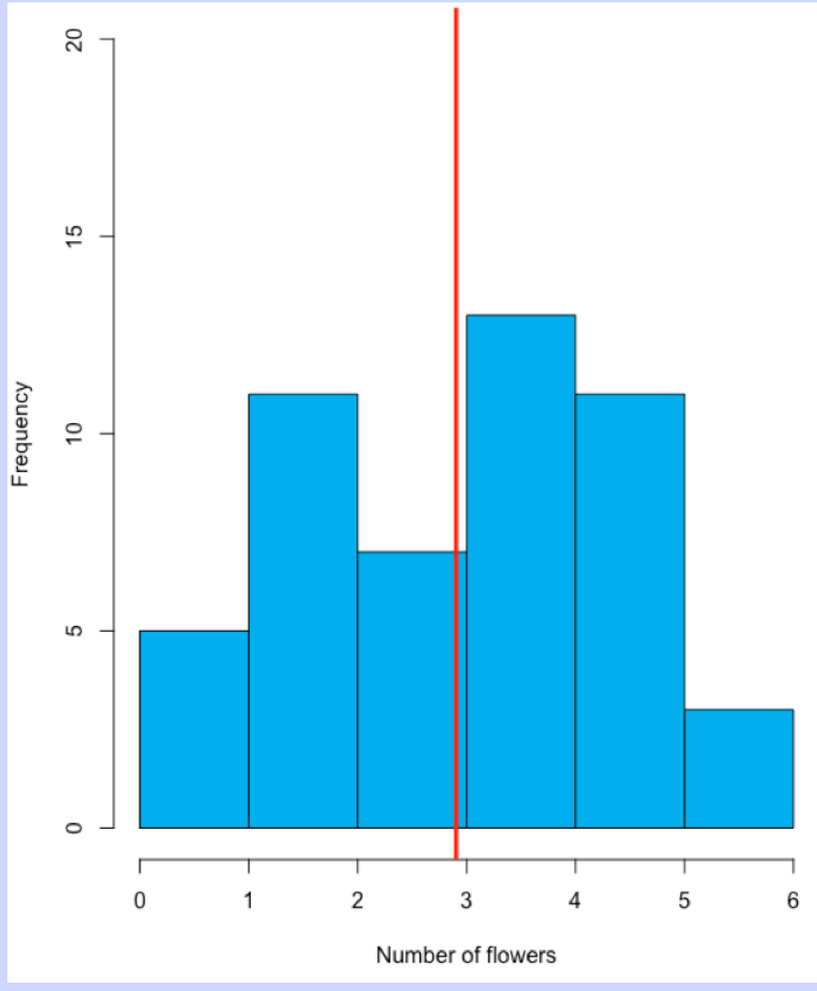
Question: Are the two groups significantly different in the number of flowers they produce?

Group A



$$\mu = 4.03$$

Group B



$$\mu = 2.90$$

T-test (parametric)

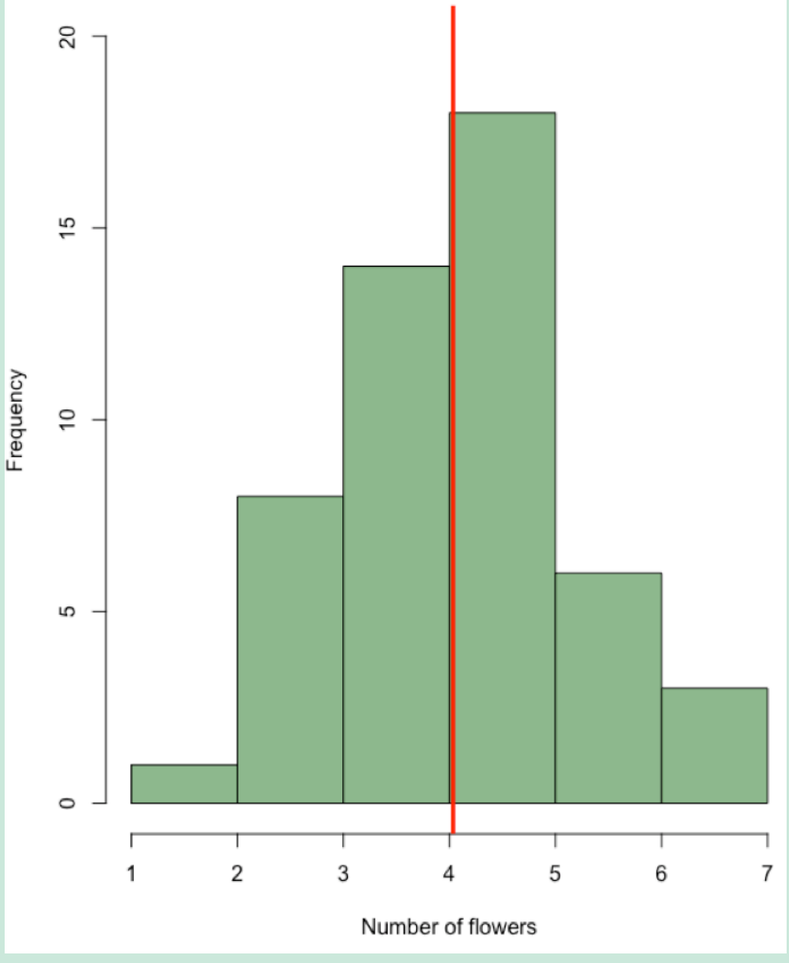
- The data is collected from a representative, **randomly** selected portion of the total population.
- The data follow a **normal distribution**.
- A reasonably **large sample size** is used (N > 30).
- **Homogeneity** of variance.

Test of normality

Shapiro test: The null-hypothesis of this test is that the population is normally distributed. Thus, on the one hand, if the p-value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are **not normally distributed**.

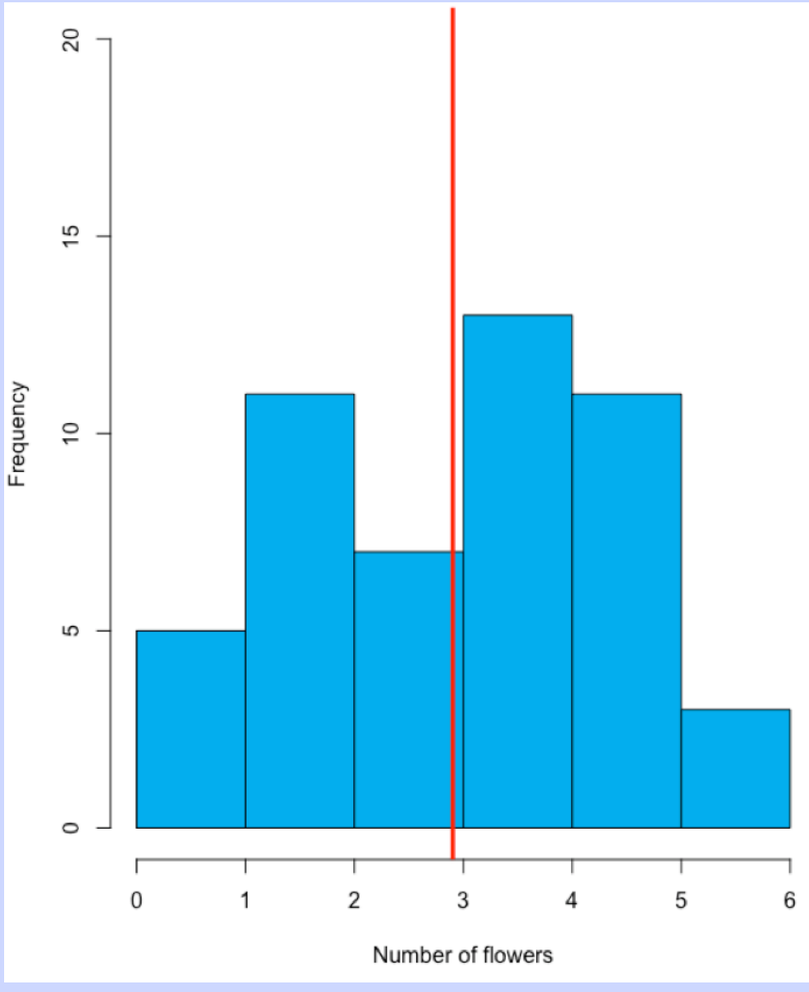
```
> shapiro.test(groupA)
```

Group A



$$P-value = 0.34 \text{ (n.s.)}$$

Group B



$$P-value = 0.27 \text{ (n.s.)}$$

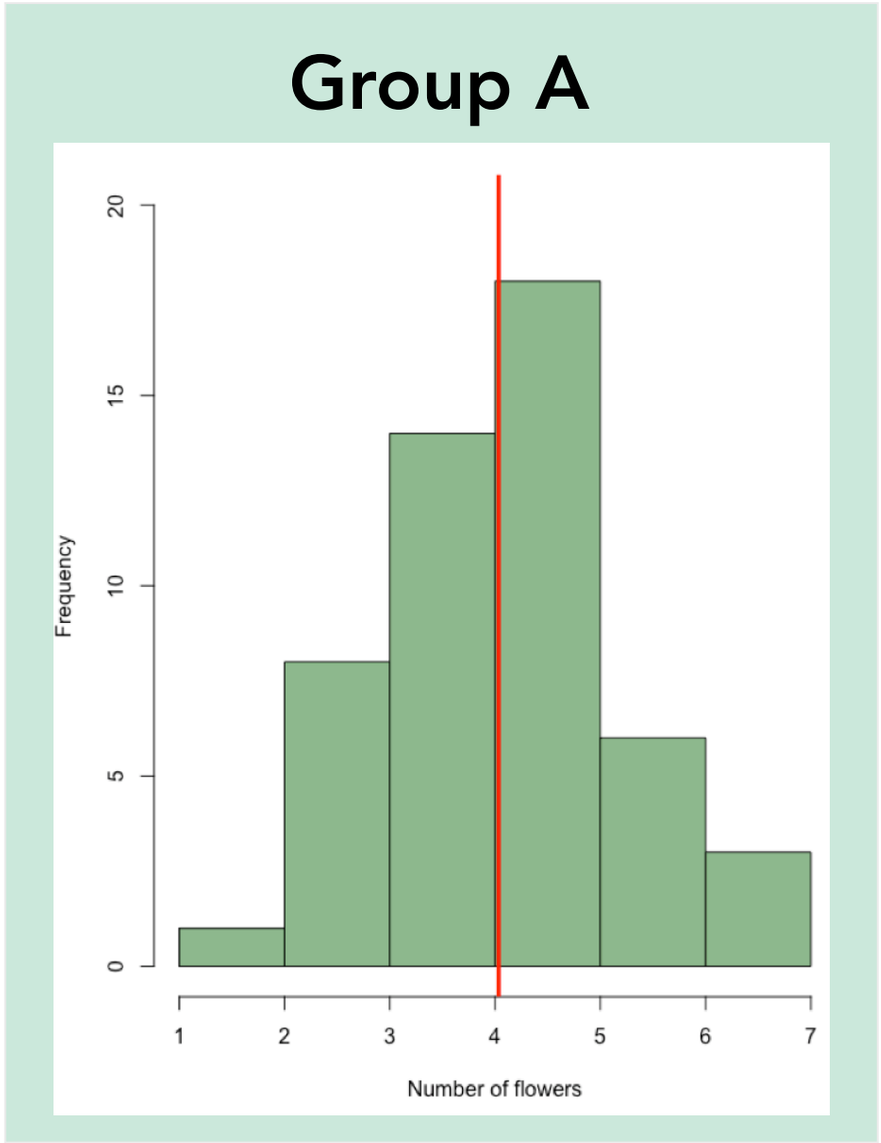
P-value > 0.05 → the data of both samples does not deviate from a normal distribution.

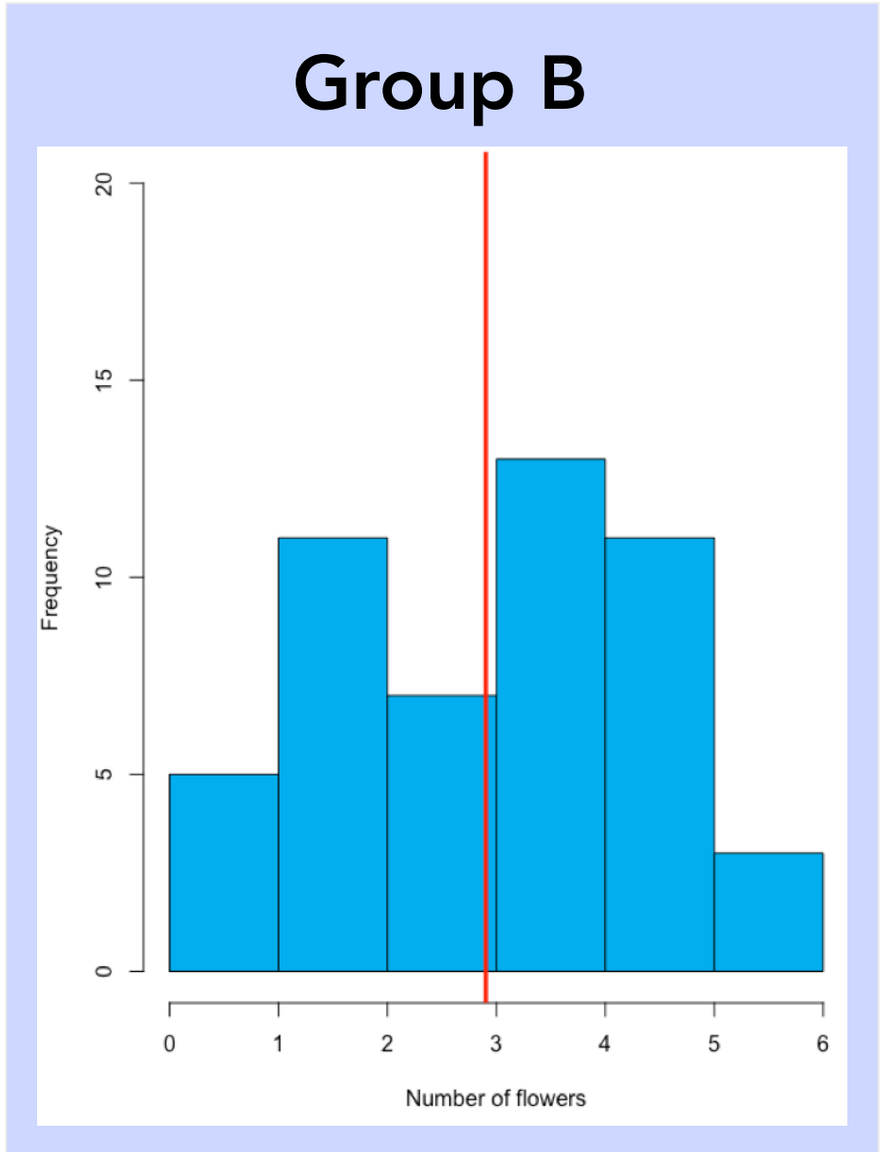
Homogeneity of variance

Bartlett's test: In statistics, Bartlett's test is used to test if *k* samples are from populations with equal variances.

Bartlett's test is **sensitive to departures from normality**. Levene's test and the Brown–Forsythe test are alternatives to the Bartlett test that are less sensitive to departures from normality.

```
> barlett.test(groupA, groupB)
```





$$P-value = 0.078 \text{ (n.s.)}$$

Bartlett's test is used to **test** the **null hypothesis**, H_0 that all k population variances are equal against the alternative that at least two are different.

T-test (parametric)

- ☑ The data is collected from a representative, **randomly** selected portion of the total population.
- ☑ The data follow a **normal distribution**.
- ☑ A reasonably **large sample size** is used (N > 30).
- ☑ **Homogeneity** of variance.

Question: Are the two Groups significantly different in the number of flowers they produce?

```
t.test(groupA, groupB, var.equal = TRUE)

t = 4.3915, df = 98, p-value = 2.848e-05

95 percent confidence interval:
 0.6198587 1.6419286

mean in group A mean in group B
       4.033558        2.902665
```

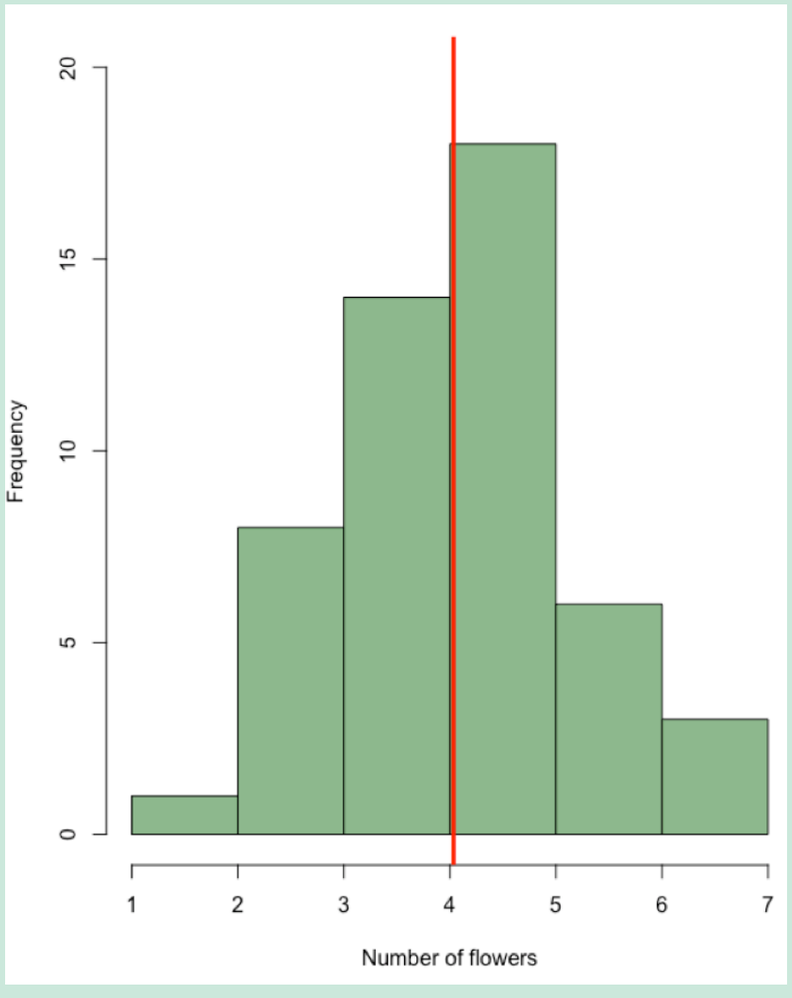
Answer: **The two Groups are significantly different in the number of flowers they produce.**

t value

The calculations behind t-values compare your sample mean(s) to the null hypothesis and incorporates both the sample size and the variability in the data. A t-value of 0 indicates that the sample results exactly equal the null hypothesis. As the difference between the sample data and the null hypothesis increases, the absolute value of the t-value increases.

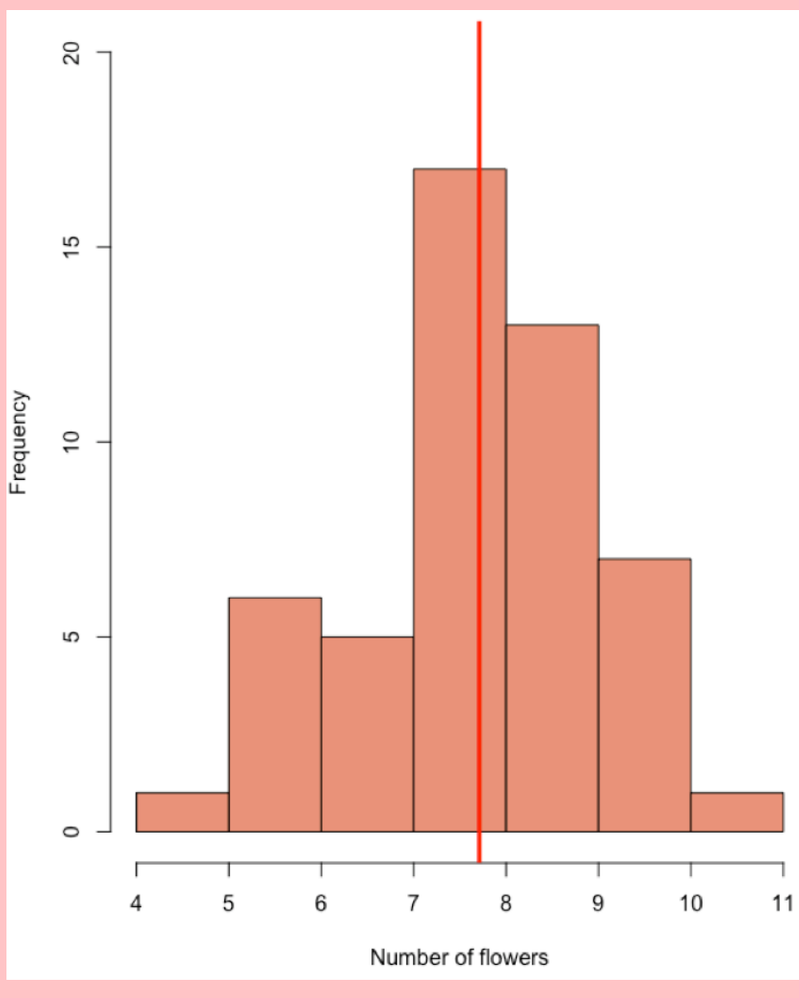
Question: Are the three groups significantly different in the number of flowers they produce?

Group A



$\mu = 4.03$

Group C



$\mu = 7.71$

t-value

The calculations behind t-values compare your sample mean(s) to the null hypothesis and incorporates both the sample size and the variability in the data. A t-value of 0 indicates that the sample results exactly equal the null hypothesis. As the difference between the sample data and the null hypothesis increases, the absolute value of the t-value increases.

```
t.test(groupA, groupB)

t = -4.3915, df = 98,
p-value = 2.848e-05

mean of x mean of y
 2.902665  4.033558
```

```
t.test(groupC, groupA)

t = 14.9, df = 98, p-
value < 2.2e-16

mean of x mean of y
 7.710984  4.033558
```

Table entries are values of t corresponding to proportions in one tail or in two tails combined.

	Proportion in One Tail					
	0.25	0.10	0.05	0.025	0.01	0.005
	Proportion in Two Tails Combined					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

Table III of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed. London: Longman Group Ltd., 1974 (previously published by Oliver and Boyd Ltd., Edinburgh). Adapted and reprinted with permission of the Addison Wesley Longman Publishing Co.

```
t.test(groupA, groupB)

t = -4.3915, df = 98,
p-value = 2.848e-05

mean of x mean of y
 2.902665  4.033558
```

```
t.test(groupC, groupA)

t = 14.9, df = 98, p-
value < 2.2e-16

mean of x mean of y
 7.710984  4.033558
```

Important to remember!

We noted previously that one of the assumptions for the t-test is that the **variances of the two samples are equal**. However, a modification of the t-test known as Welch's test is said to correct for this problem by estimating the variances, and adjusting the degrees of freedom to use in the test.

```
t.test(sampleA, SampleB, var.equal = TRUE)
```

```
## Default S3 method:
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Important to remember!

```
t.test(groupB, groupA, var.equal = FALSE)
t = -4.3915, df = 92.284, p-value = 2.996e-05
```

```
t.test(groupB, groupA, var.equal = TRUE)
t = -4.3915, df = 98, p-value = 2.848e-05
```

Important to remember!

If you want to measure a trait before and after treatment (eg. effect of nutrient on plant growth), you need to used a **paired sample T-test** since the measurements are not independent.

```
t.test(sampleA, SampleB, paired = TRUE)
```

```
## Default S3 method:
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Important to remember!

```
t.test(groupB, groupA, paired = FALSE)
t = -4.3915, df = 92.284, p-value = 2.996e-05
```

```
t.test(groupB, groupA,  paired = TRUE)
t = -4.6764, df = 49, p-value = 2.324e-05
```

What is the degree of freedom (df)?

degree of freedom

The **t.test( )** function produces a variety of t-tests. Unlike most statistical packages, the default assumes unequal variance and applies the Welsh df modification.

```
t.test(groupA, groupB, var.equal = FALSE)
t.test(groupA, groupB, var.equal = TRUE)
```

degree of freedom

general definition - the number of **independently variable factors** affecting the range of states in which a system may exist.

in statistics - the number of **independent values or quantities** that can be assigned to a **statistical distribution**.

5 choices of flavours

5 choices of flavours

but the freedom of choosing a flavour is 4


MORE
THINGS
CONSIDERED

One Sample t-Test

The One Sample t-Test determines whether the sample mean is statistically different from a known or hypothesised population mean.

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \text{ where } s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

```
sample <- 1:9
mean(sample)
t.test(sample, mu=5, conf.level = 0.95)
# or
sample <- (rnorm(1000, mean = 0, sd = 0.5))
t.test(sample, mu = 0, conf.level = 0.95)
```

1	2	3	4	5	6	7	8	9	10	mean
5	5	5	5	5	5	5	5	5	**5**	5
6	4	7	3	8	2	9	1	5	**5**	5
1	1	1	1	1	1	1	1	1	**41**	5

For a one-sample t-test, one degree of freedom is spent estimating the mean, and the remaining ***n* - 1 degrees** of freedom estimate variability.

Chi-Square Test of Independence

A chi-square test of independence is used to determine whether two categorical variables are dependent.

	A1	A2	
B1	6	4	10
B2	8	6	14
	14	10	24

	A1	A2	
B1	**6**		10
B2			14
	14	10	24

d.f. = 1

Chi-Square Test of Independence

	A1	A2	A3	
B1				20
B2				20
B3				20
	20	30	10	60

How much freedom is there in a 3x3 table?

Chi-Square Test of Independence

	A1	A2	A3	
B1	10	0	10	20
B2	5	5	10	20
B3	5	5	10	20
	20	10	30	60

	A1	A2	A3	
B1	5	0	15	20
B2	10	10	0	20
B3	5	0	15	20
	20	10	30	60

	A1	A2	A3	
B1				20
B2				20
B3				20
	20	10	30	60

Chi-Square Test of Independence

	A1	A2	A3	
B1	10	0	10	20
B2	5	5	10	20
B3	5	5	10	20
	20	10	30	60

	A1	A2	A3	
B1	**10**	**0**	10	20
B2	5	**5**	10	20
B3	5	5	**10**	20
	20	10	30	60

	A1	A2	A3	
B1	**10**	0	10	20
B2	**5**	**5**	10	20
B3	**5**	5	10	20
	20	10	30	60

d.f. = 4

Chi-Square Test of Independence

	A1	A2	A3	A4	A5	
B1						66
B2						40
B3						32
B4						57
	23	35	33	44	60	

How much freedom is there in a 4x6 table?

Chi-Square Test of Independence

	A1	A2	A3	A4	A5	
B1	12	14	3	17	20	66
B2	5	6	10	6	13	40
B3	5	8	9	6	4	32
B4	1	7	11	15	23	57
	23	35	33	44	60	

d.f. = (r-1)(c-1)=12

Testing Hardy-Weinberg

One approach to testing the hypothesis that genotypes are in Hardy-Weinberg proportions is the $\chi 2$ for goodness of fit between observed and predicted genotype (or phenotype) frequencies.

We have to know how many degrees of freedom are associated with the test.

How many degrees of freedom does a chi square test have? Normally, the degrees of freedom are equal to the number of observations minus 1. However, this is not true for testing the Hardy-Weinberg equilibrium. Instead, you have to think about how many quantities are really free to vary. Remember that we used the population to estimate p and q, then we used p and q to get p^2, 2pq, and q^2. Once we decided on a value for p, everything else was decided for us. Whatever p was, q had to be 1-p, and p2, 2pq, and q2 were set as well.

	A1	A2	A3	
A1	10	5	10	25
A2	5	5	0	10
A3	10	0	10	20
	25	10	20	55

	A1	**A2**	**A3**	
A1	10	5	10	25
A2	5	5	0	10
A3	10	0	10	20
	25	10	20	55

	A1	A2	A3	
A1	**10**			25
A2		**5**		10
A3		**5**	**10**	20
	25	10	20	55

	A1	**A2**	**A3**	
A1	10	5	10	25
A2	5	5	0	10
A3	10	0	10	20
	25	10	20	55

	A1	A2	A3	
A1	**10**	0	5	25
A2	0	**5**	5	10
A3	5	5	**10**	20
	25	10	20	55

	A1	**A2**	**A3**	
A1	10	5	10	25
A2	5	5	0	10
A3	10	0	10	20
	25	10	20	55

	A1	A2	A3	
A1	**10**	0	5	25
A2	0	**5**	5	10
A3	5	5	**10**	20
	25	10	20	55

	A1	A2	A3	
A1	**10**			25
A2		**5**		10
A3		**4**	**10**	20
	25	10	20	55

	A1	**A2**	**A3**	
A1	10	5	10	25
A2	5	5	0	10
A3	10	0	10	20
	25	10	20	55

	A1	A2	A3	
A1	**10**	0	5	25
A2	0	**5**	5	10
A3	5	5	**10**	20
	25	10	20	55

	A1	A2	A3	
A1	**10**	1	6	25
A2	1	**5**	4	10
A3	6	4	**10**	20
	25	10	20	55

	A1	A2	A3	
A1	**10**	2	7	25
A2	2	**5**	3	10
A3	7	3	**10**	20
	25	10	20	55

	A1	A2	A3	
A1	**10**	3	8	25
A2	3	**5**	2	10
A3	8	2	**10**	20
	25	10	20	55

	A1	A2	A3	
A1	**10**	4	9	25
A2	4	**5**	1	10
A3	9	1	**10**	20
	25	10	20	55

	A1	A2	A3	
A1	**10**	5	10	25
A2	5	**5**	0	10
A3	10	0	**10**	20
	25	10	20	55

d.f. = 3

Number of alleles: 3

$$Genotypes = \frac{n(n+1)}{2}$$

Number of genotypes: 6

$$d.f. = \#\,genotypes - \#\,alleles$$

Degree of freedom: 3

Number of alleles: 4

$$Genotypes = \frac{n(n+1)}{2}$$

Number of genotypes: 10

$$d.f. = \#\,genotypes - \#\,alleles$$

Degree of freedom: 6

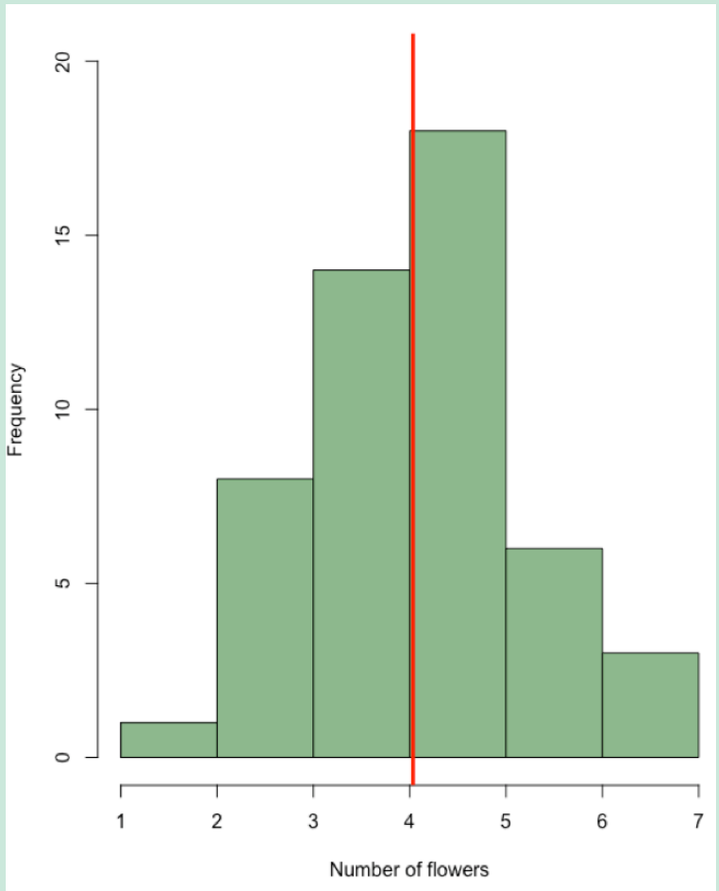
Non-parametric test

The **Mann–Whitney *U* test** (also called the Mann–Whitney–Wilcoxon, Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. **Unlike the *t*-test it does not require the assumption of normal distributions.**

```
wilcoxon.test(groupA, groupB, paired = FLASE)
wilcoxon.test(groupA, groupB, paired = TRUE)
```

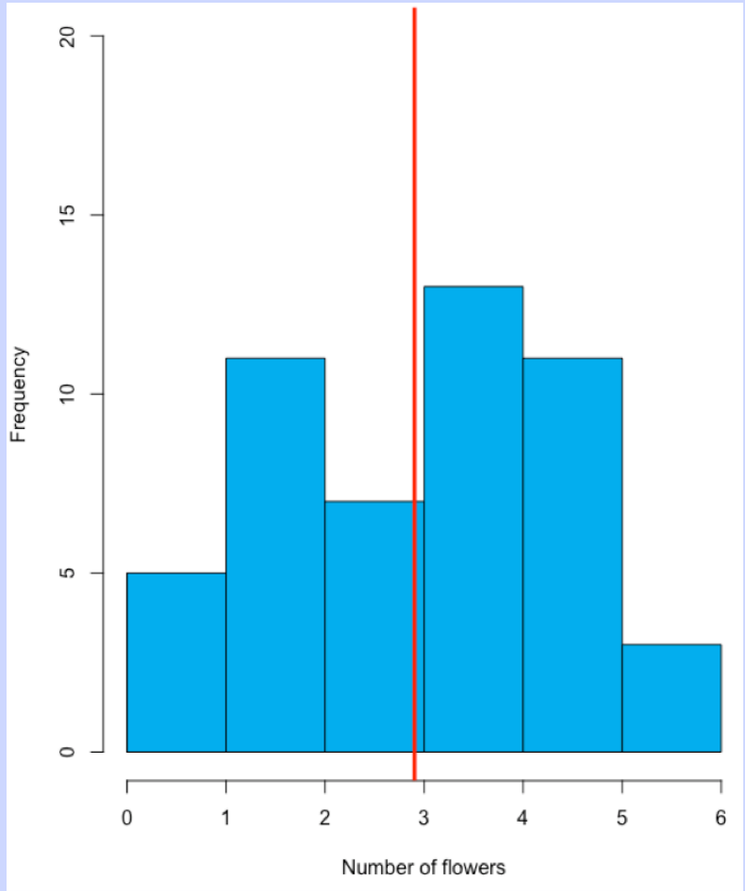
Question: Are the three groups significantly different in the number of flowers they produce?

Group A



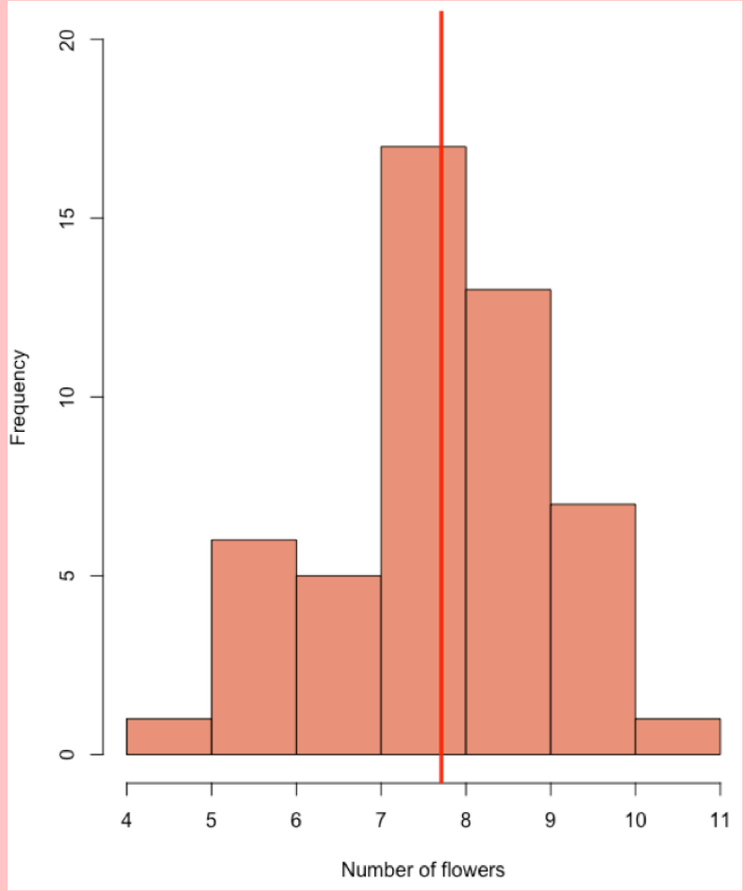
$\mu = 4.03$

Group B



$\mu = 2.90$

Group C

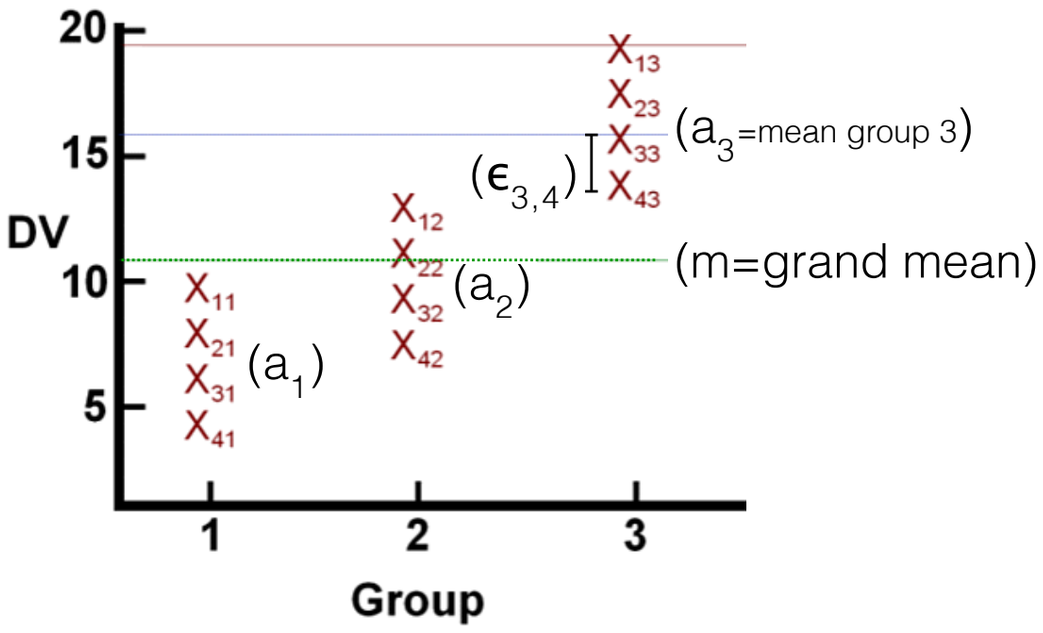


$\mu = 7.71$

Anova (parametric)

- The data is collected from a representative, **randomly** selected portion of the total population.
- The data follow a **normal distribution** (or the **residuals** of the model are normally distributed)
- A reasonably **large sample size** is used (N > 20).
- **Homogeneity** of variance.

Normality of the residuals (or errors)



$$y_{ij} = m + a_i + \epsilon_{ij}$$

$$y_{43} = m + a_3 + \epsilon_{3,4}$$

The equation indicates that the *j*th data value, from level *i*, is the sum of three components: the common value (grand mean), the level effect (the deviation of each level (or group) mean from the grand mean), and the residual (what's left over).

Normality of the residuals (or errors)

value	group
2.662924	A
3.861776	A
3.874668	A
3.601635	A
1.440844	A
2.065737	A

```
anova_model <- aov(data$value ~ data$group)
shapiro.test(residuals(anova_model))


data:  residuals(anova_model)
W = 0.99153, p-value = 0.5128
```

P-value > 0.05 → The residuals of the anova model do not deviate from a normal distribution.

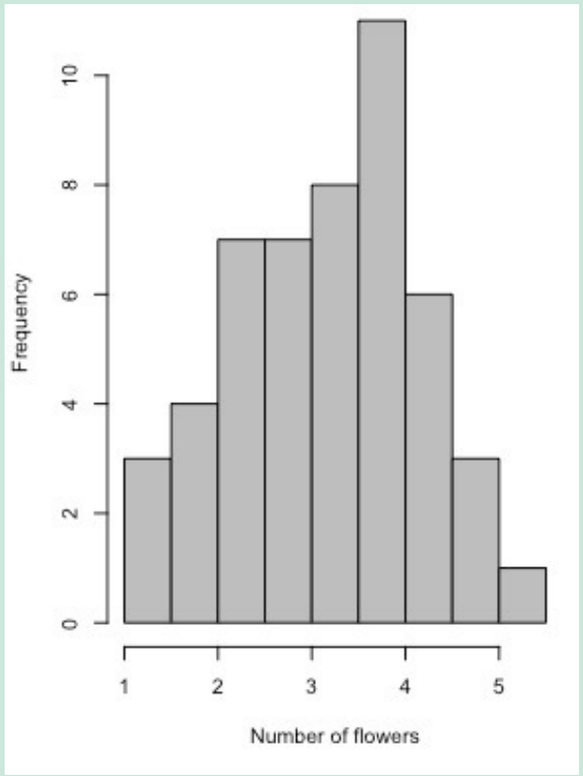
Homogeneity of variance

Performs Bartlett's test of the null that the variances in each of the groups (samples) are the same.

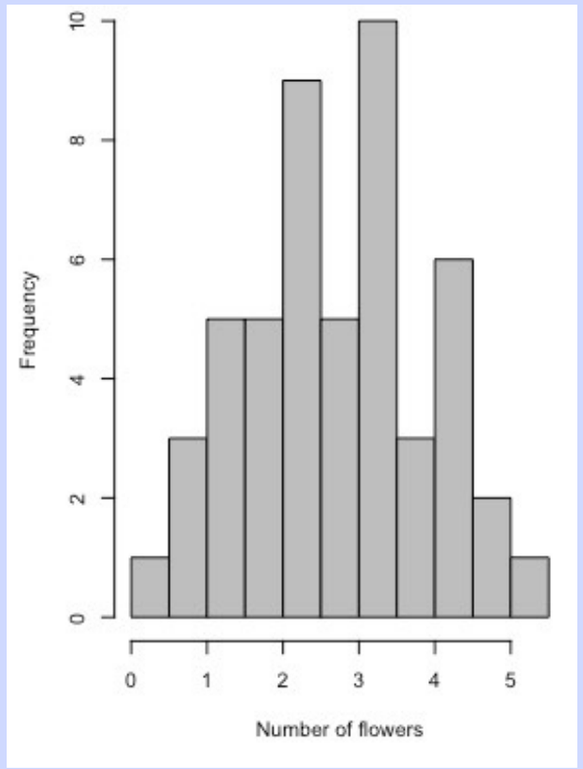
```
bartlett.test(data$value ~ data$group)
```

Homogeneity of variance

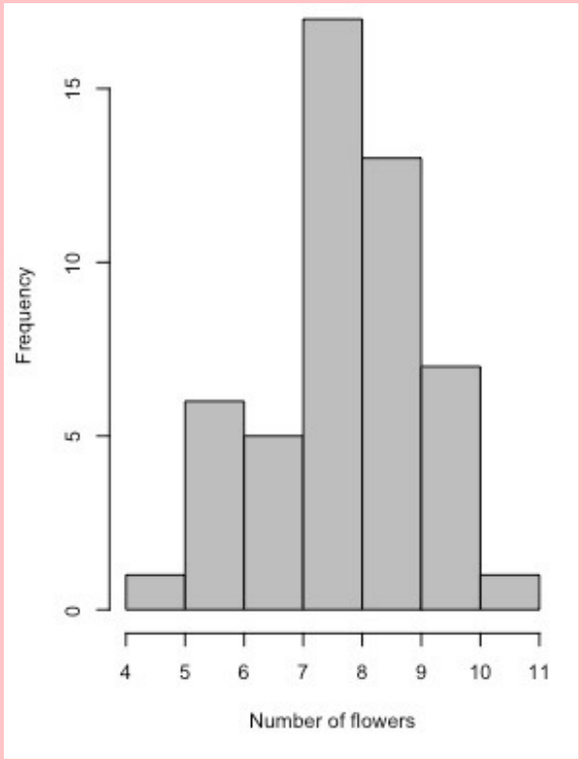
Group A



Group B



Group C




```
Bartlett's K-squared = 3.216, df = 2, p-value = 0.20
```

Anova (parametric)

- ☑ The data is collected from a representative, **randomly** selected portion of the total population.
- ☑ The data follow a **normal distribution** (or the **residuals** of the model are normally distributed)
- ☑ A reasonably **large sample size** is used (N > 20).
- ☑ **Homogeneity** of variance.

Analysis of variance (**ANOVA**) is used to analyze the differences among group means in a sample. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalizes the *t*-test to more than two groups.

```
anova_model <- aov(value ~ group, data = data)
summary(anova_model)

             Df Sum Sq Mean Sq F value Pr(>F)
group         2  632.0  316.02   185.3 <2e-16 ***
```

At this point, it is important to realize that the ANOVA **cannot tell you which specific group(s)** were statistically significantly different from each other, only that at least two groups were.

To determine which groups differed from each other, you need to use a **post hoc test** (e.g. Tukey's range test).

Tukey's range test, also known as the Tukey's test, Tukey method, Tukey's honest significance test, or Tukey's HSD is a single-step multiple comparison procedure and statistical test.

It can be used on raw data or in conjunction with an ANOVA (post-hoc analysis) to find means that are significantly different from each other.

```
TukeyHSD(anova_model)

$group
          diff        lwr         upr     p adj
B-A -1.130894 -1.749297 -0.5124904 8.12e-05
C-A  3.677426  3.059023  4.2958293 0.00e+00
C-B  4.808320  4.189916  5.4267229 0.00e+00
```

Groups A and B are not significantly different

Group C is significantly different from A and B

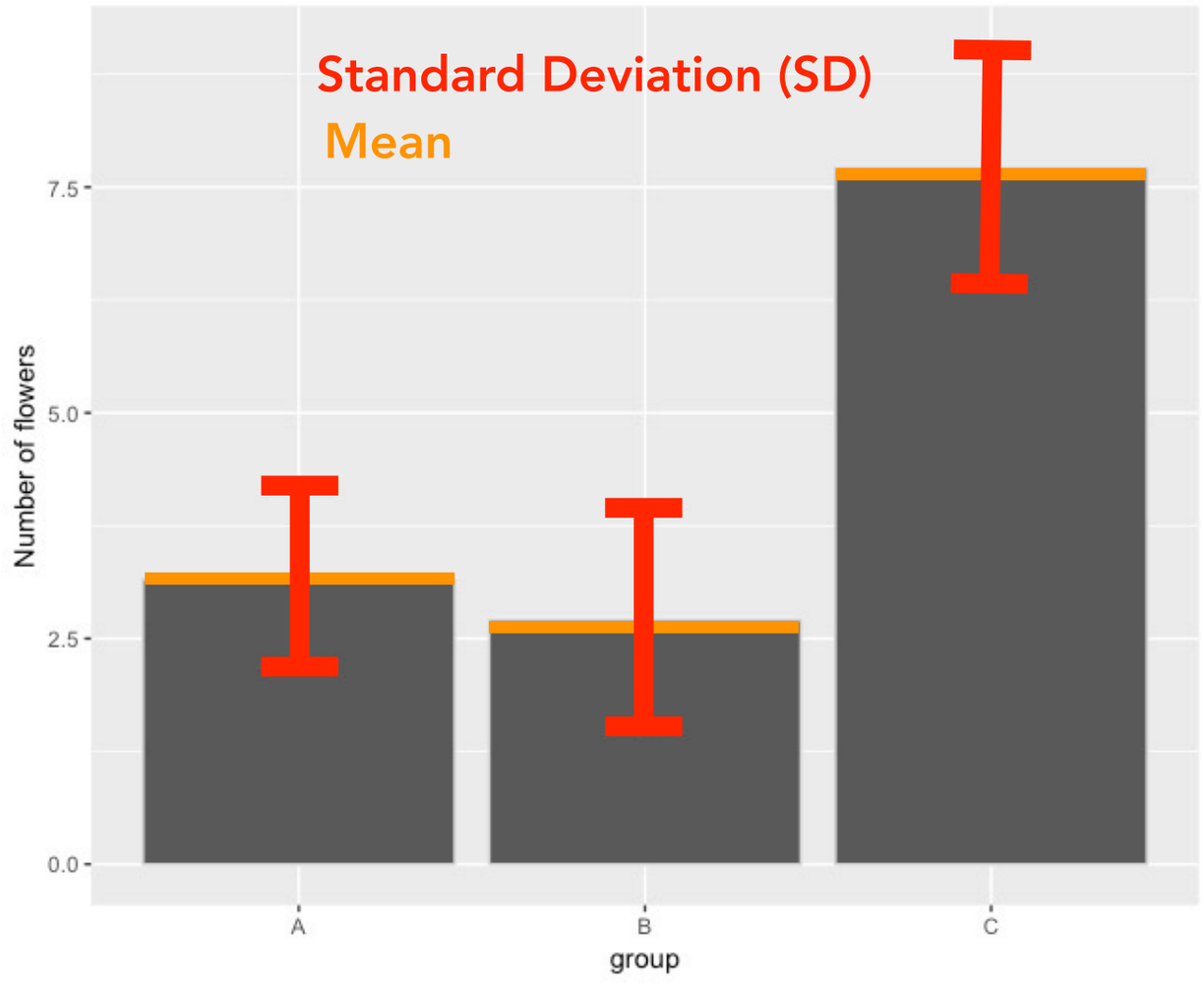
Non-parametric test

The **Kruskal–Wallis test** or **one-way ANOVA on ranks** is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes. It extends the Mann–Whitney *U* test, which is used for comparing only two groups.

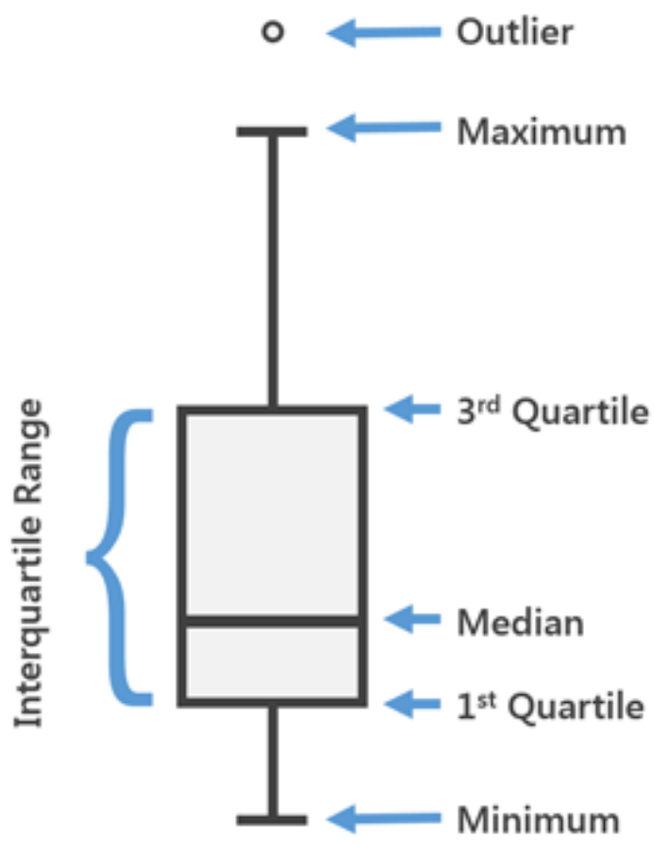
```
kruskal_model <- kruskal.test(value ~ group, data = data)

data:  value by group
Kruskal-Wallis chi-squared = 100.19, df = 2, p-value < 2.2e-16
```

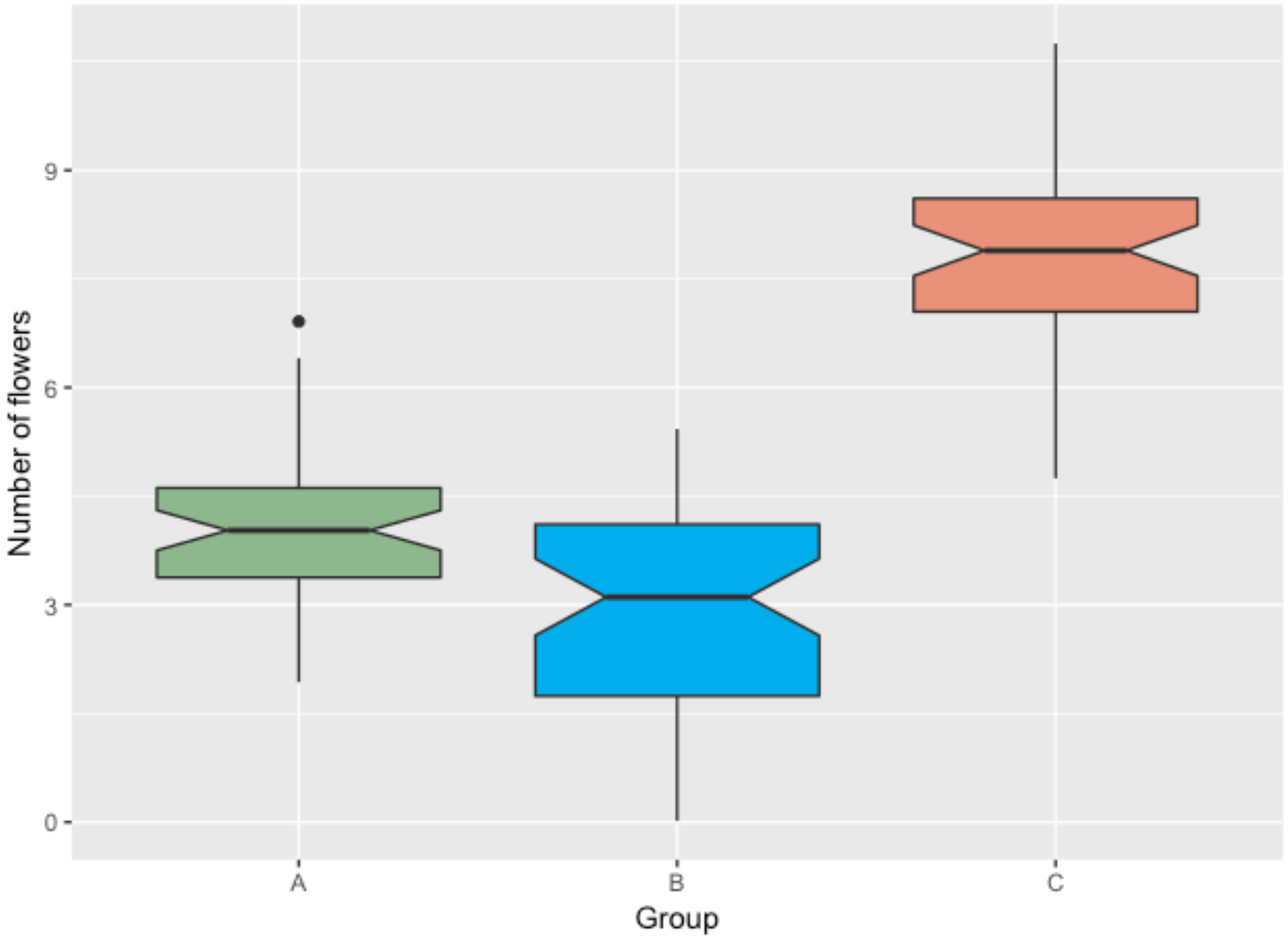
Did you know that barplot are not he only way to visualise your data?



What about using **boxplots** instead?



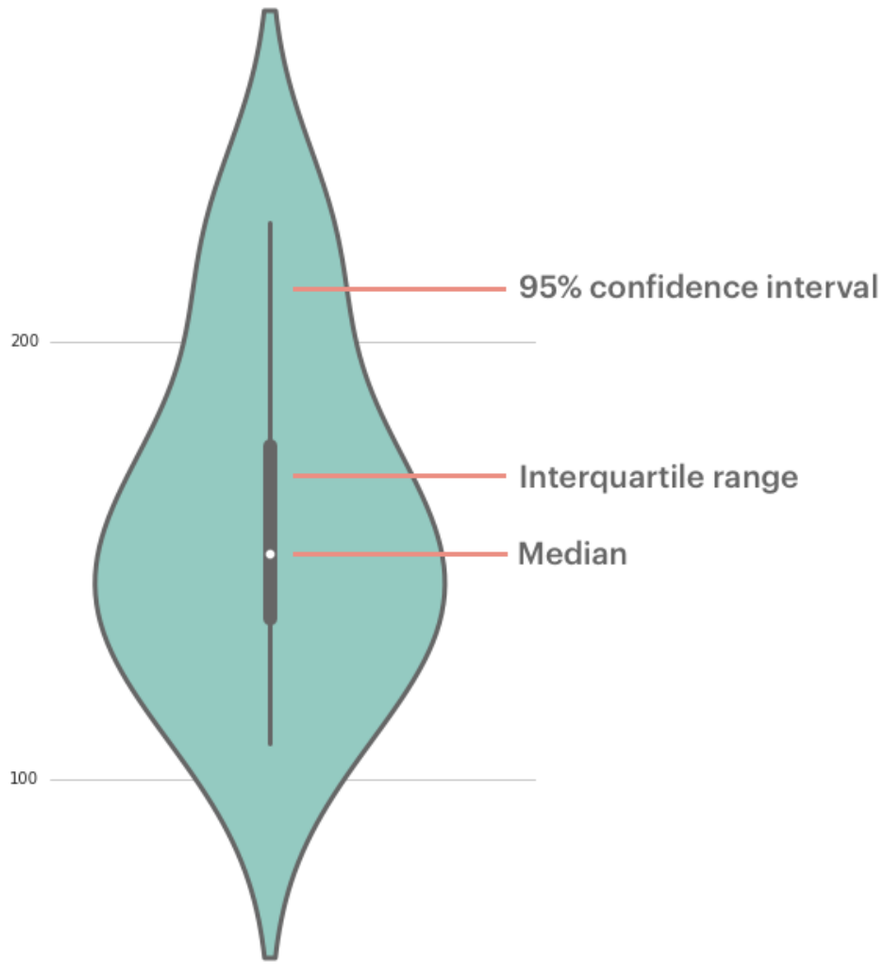
Boxplot



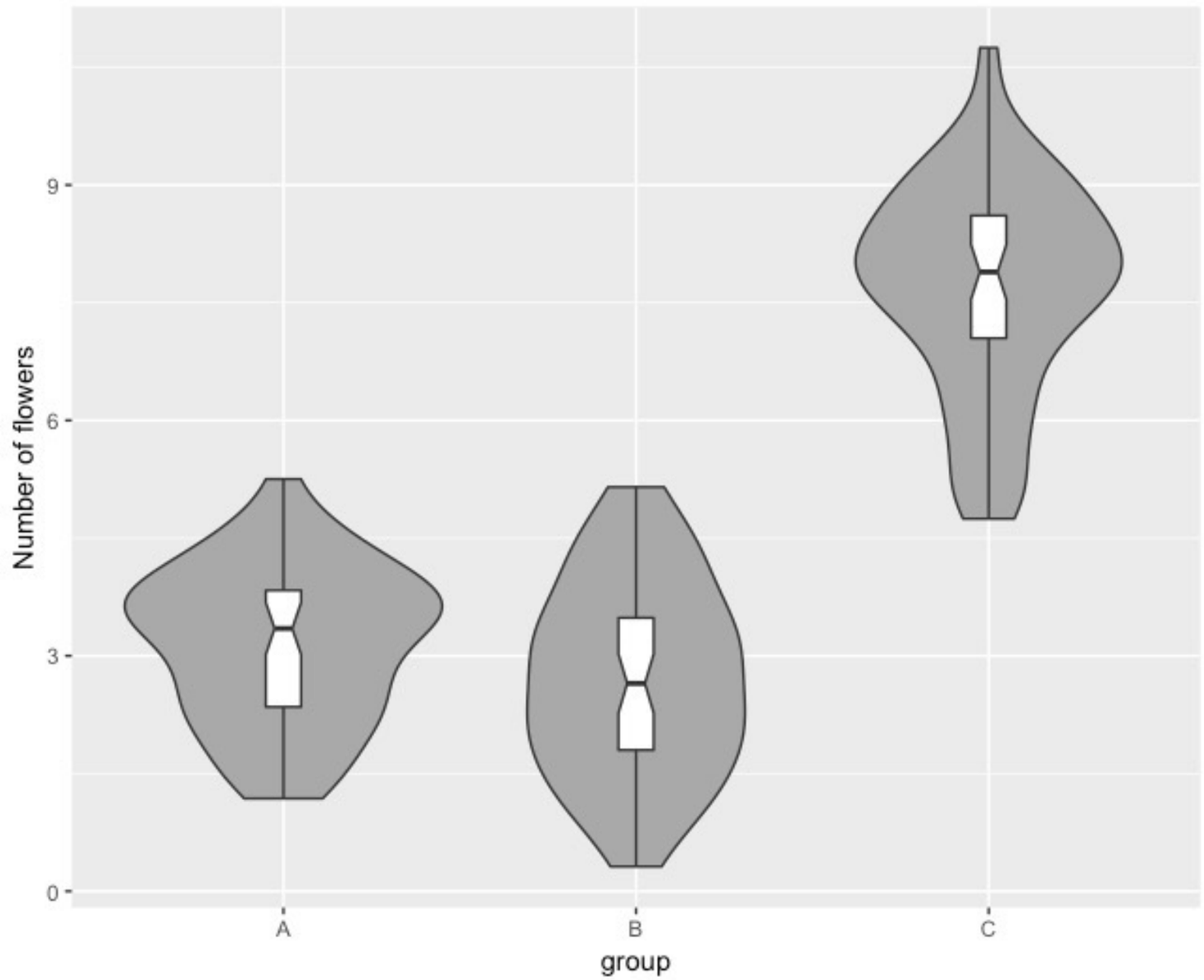
You can confirm visually the results of the ANOVA:

- Group A and B **notches overlap**
- Group C and A/B **notches do not overleap**

Violin Plot


95% confidence interval
200
Interquartile range
Median
100

Violin Plot



Violin plots are similar to **box plots**, except that they also show the probability density of the data at different values.

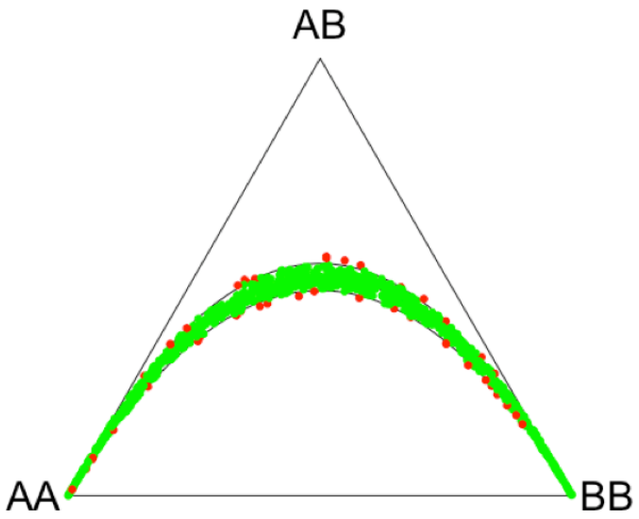
Violin plots are especially informative for large dataset.

Genetic Diversity Centre Zurich

Multiple Comparisons

Genetic Diversity Centre Zurich

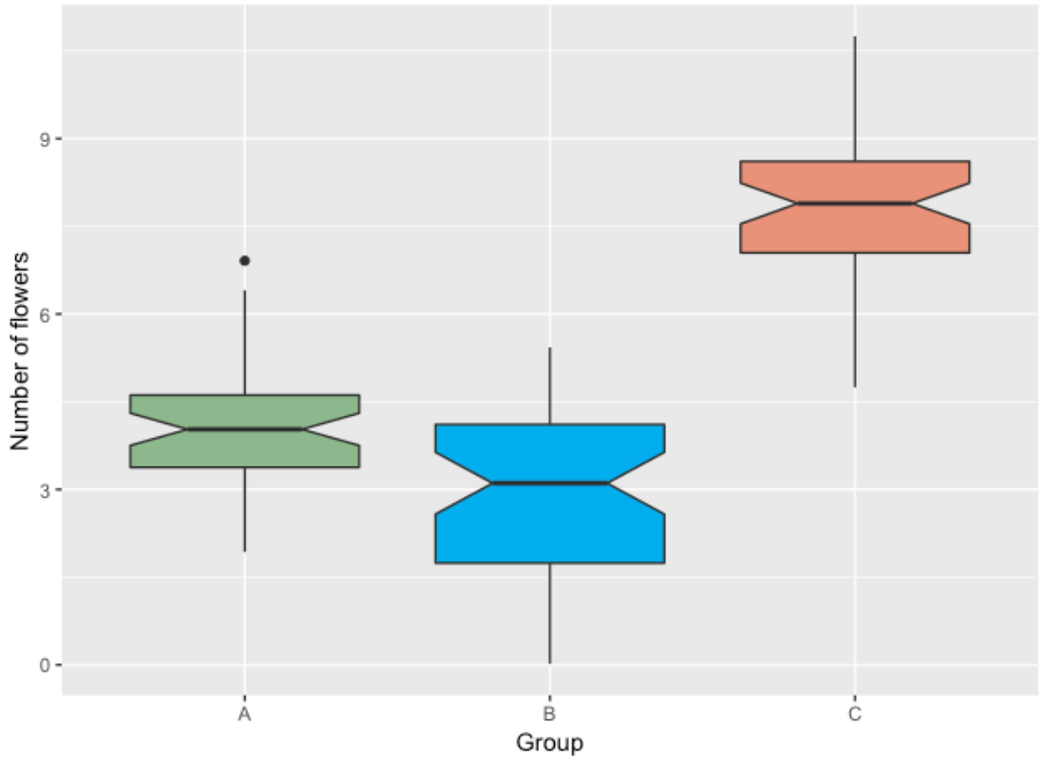
The problem - When you perform a large number of statistical tests, some will have *P*-values less than 0.05 purely by chance, even if all your null hypotheses are true. The **Bonferroni** correction is one simple way to take this into account. Adjusting the false discovery rate using the **Benjamini-Hochberg** procedure is a more powerful method.




```
library(HardyWeinberg)
n   <- 1000          # sample size
m   <- 1000          # number of markers
out <- HWData(n,m) # create a random data set
HWTernaryPlot(out, 1000, region=1,
   hwcurve=TRUE, vbounds=FALSE, vertex.cex=2)
```

The problem of **multiple comparisons** (multiple statistical tests) has received increasing attention in the last few years. This is important for such techniques as the use of RNASeq quantities or evolutionary genomics, where the sequences of every gene in the genome of two or more species can be compared. There is no universally accepted approach for dealing with the problem; it is an area of active research.

The classic approach to the multiple comparison problem is to control for the error rate. Instead of setting the critical P level for significance to e.g. 0.05, you use a adjusted lower threshold. The most common adjustment method is the **Bonferroni correction**. The idea is simple, if you are doing 100 statistical tests, the critical value for an individual test would be 0.05/100 = 0.0005, and you would only consider individual tests with P-vlaues < 0.0005 (instead of 0.05) to be significant.



P-value: 0.05

Bonferroni adjusted P-value: 0.05/3 = 0.0167

```
t.test(groupB, groupA, var.equal = TRUE)

t = -4.3915, df = 98, p-value = 2.848e-05
```

```
t.test(groupC, groupA, var.equal = TRUE)

t = 14.9, df = 98, p-value < 2.2e-16
```

```
t.test(groupB, groupA, var.equal = TRUE)

t = 17.281, df = 98, p-value < 2.2e-16
```

The Bonferroni correction is appropriate when a single false positive in a set of tests would be a problem. It is mainly useful when there are a fairly small number of multiple comparisons and you're looking for one or two that might be significant. However, if you have a large number of multiple comparisons and you're looking for many that might be significant, the Bonferroni correction may lead to a very high rate of **false negatives**.

Calorie intake, olive oil consumption and mammographic density among Spanish women

Nicolás García-Arenzana[1,2], Eva María Navarrete-Muñoz[3,4], Virginia Lope[1,4], Pilar Moreo[5], Carmen Vidal[6], Soledad Laso-Pablos[7,8], Nieves Ascunce[4,9], Francisco Casanova-Gómez[10], Carmen Sánchez-Contador[11], Carmen Santamariña[12], Nuria Aragonés[1,4], Beatriz Pérez Gómez[1,4], Jesús Vioque[3,4] and Marina Pollán[1,4]

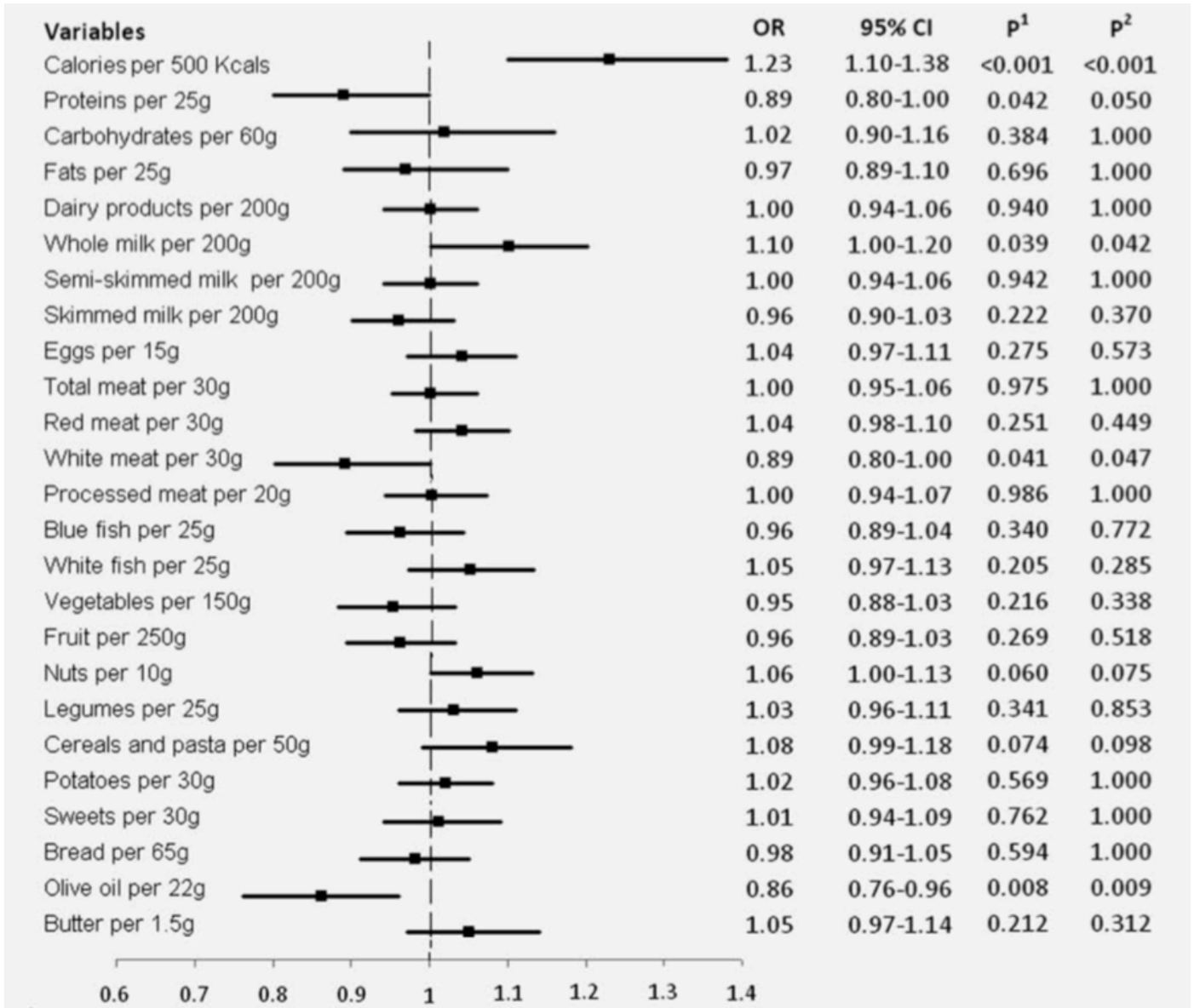
Dietary variable	*P* value	Rank	(i/m)Q
Total calories	<0.001	1	0.010
Olive oil	0.008	2	0.020
Whole milk	0.039	3	0.030
White meat	0.041	4	0.040
Proteins	0.042	5	0.050
Nuts	0.060	6	0.060
Cereals and pasta	0.074	7	0.070
White fish	0.205	8	0.080
Butter	0.212	9	0.090
Vegetables	0.216	10	0.100
Skimmed milk	0.222	11	0.110
Red meat	0.251	12	0.120
Fruit	0.269	13	0.130
Eggs	0.275	14	0.140
Blue fish	0.34	15	0.150
Legumes	0.341	16	0.160
Carbohydrates	0.384	17	0.170
Potatoes	0.569	18	0.180
Bread	0.594	19	0.190
Fats	0.696	20	0.200
Sweets	0.762	21	0.210
Dairy products	0.94	22	0.220
Semi-skimmed milk	0.942	23	0.230
Total meat	0.975	24	0.240
Processed meat	0.986	25	0.250

One good technique for controlling the false discovery rate was briefly mentioned by Simes (1986) and developed in detail by **Benjamini and Hochberg** (1995). Put the individual *P* values in order, from smallest to largest. The smallest *P* value has a rank of *i*=1, then next smallest has *i*=2, etc. Compare each individual *P* value to its Benjamini-Hochberg critical value, (*i*/*m*)*Q*, where *i* is the rank, *m* is the total number of tests, and *Q* is the false discovery rate you choose. The largest *P* value that has *P*<(i/m)Q is significant, and *all* of the *P* values smaller than it are also significant, even the ones that aren't less than their Benjamini-Hochberg critical value.

Sources: Mc Donald (2014) Handbook of Biological Statistics
Mangiafico (2015) An R Companion for the Handbook of Biological Statistics

Calorie intake, olive oil consumption and mammographic density among Spanish women

Nicolás García-Arenzana[1,2], Eva María Navarrete-Muñoz[3,4], Virginia Lope[1,4], Pilar Moreo[5], Carmen Vidal[6], Soledad Laso-Pablos[7,8], Nieves Ascunce[4,9], Francisco Casanova-Gómez[10], Carmen Sánchez-Contador[11], Carmen Santamariña[12], Nuria Aragonés[1,4], Beatriz Pérez Gómez[1,4], Jesús Vioque[3,4] and Marina Pollán[1,4]



Variables	OR	95% CI	P¹	P²
Calories per 500 Kcals	1.23	1.10-1.38	<0.001	<0.001
Proteins per 25g	0.89	0.80-1.00	0.042	0.050
Carbohydrates per 60g	1.02	0.90-1.16	0.384	1.000
Fats per 25g	0.97	0.89-1.10	0.696	1.000
Dairy products per 200g	1.00	0.94-1.06	0.940	1.000
Whole milk per 200g	1.10	1.00-1.20	0.039	0.042
Semi-skimmed milk per 200g	1.00	0.94-1.06	0.942	1.000
Skimmed milk per 200g	0.96	0.90-1.03	0.222	0.370
Eggs per 15g	1.04	0.97-1.11	0.275	0.573
Total meat per 30g	1.00	0.95-1.06	0.975	1.000
Red meat per 30g	1.04	0.98-1.10	0.251	0.449
White meat per 30g	0.89	0.80-1.00	0.041	0.047
Processed meat per 20g	1.00	0.94-1.07	0.986	1.000
Blue fish per 25g	0.96	0.89-1.04	0.340	0.772
White fish per 25g	1.05	0.97-1.13	0.205	0.285
Vegetables per 150g	0.95	0.88-1.03	0.216	0.338
Fruit per 250g	0.96	0.89-1.03	0.269	0.518
Nuts per 10g	1.06	1.00-1.13	0.060	0.075
Legumes per 25g	1.03	0.96-1.11	0.341	0.853
Cereals and pasta per 50g	1.08	0.99-1.18	0.074	0.098
Potatoes per 30g	1.02	0.96-1.08	0.569	1.000
Sweets per 30g	1.01	0.94-1.09	0.762	1.000
Bread per 65g	0.98	0.91-1.05	0.594	1.000
Olive oil per 22g	0.86	0.76-0.96	0.008	0.009
Butter per 1.5g	1.05	0.97-1.14	0.212	0.312

The Bonferroni correction and Benjamini-Hochberg procedure assume that the individual tests are **independent** of each other. That might not be the case and to account for this you might consider **multivariate analysis of variance** (manova).

Reiner, A., D. Yekutieli and Y. Benjamini. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19: 368-375.

Independent Random Sampling: MANOVA assumes that the observations are independent of one another, there is not any pattern for the selection of the sample, and that the sample is completely random.

Level and Measurement of the Variables: MANOVA assumes that the independent variables are categorical and the dependent variables are continuous or scale variables.

Absence of multicollinearity: The dependent variables cannot be too correlated to each other.

Normality: Multivariate normality (Shapiro-Wilk test) is present in the data.

Homogeneity of Variance: Variance between groups is equal.

Levene's Test of Equality of Variance: Used to examine whether or not the variance between independent variable groups are equal; also known as homogeneity of variance Non-significant values of Levene's test indicate equal variance between groups.

Box's M Test: Used to know the equality of covariance between the groups. This is the equivalent of a multivariate homogeneity of variance. Usually, significance for this test is determined at $\alpha = .001$ because this test is considered highly sensitive.

Partial eta square: Partial eta square ($\eta 2$) shows how much variance is explained by the independent variable. It is used as the effect size for the MANOVA model.

Post hoc test: If there is a significant difference between groups, then post hoc tests are performed to determine where the significant differences lie (i.e., which specific independent variable level significantly differs from another).

Multivariate F-statistics: The F-statistic is derived by essentially dividing the means sum of the square (SS) for the source variable by the source variable mean error (ME or MSE).

```
## Dataset (for help use: ? iris)
iris

## Subset data
sepl <- iris$Sepal.Length
petl <- iris$Petal.Length

## Shapiro-Wilk normality test
# Prepare data
iris.t <- t(iris[,c(1,3)])
# Run test
mshapiro.test(iris.t)

## MANOVA test
res.man <- manova(cbind(Sepal.Length,
Petal.Length) ~ Species, data = iris)
summary(res.man)
## Which differ?
summary.aov(res.man)
```

The methods Holm, Hochberg, Hommel, and Bonferroni control the family-wise error rate. These methods attempt to limit the probability of even one false discovery (a type I error, incorrectly rejecting the null hypothesis when there is no real effect), and so are all relatively strong (conservative).

The methods BH (Benjamini–Hochberg, which is the same as FDR in R) and BY control the false discovery rate. These methods attempt to control the expected proportion of false discoveries.

```
Data$Bonferroni <-
      p.adjust(data$raw.p,
               method = "bonferroni")

Data$BH <-
      signif(p.adjust(data$Raw.p,
               method = "BH"), 4)

Data$Holm <-
      p.adjust(data$raw.p,
               method = "holm")

Data$Hochberg <-
      p.adjust(data$raw.p,
               method = "hochberg")

Data$Hommel <-
      p.adjust(data$raw.p,
               method = "hommel")

Data$BY <-
      signif(p.adjust(data$ raw.p,
               method = "BY"), 4)
```

Bayesian statistics

The Posterior	The Evidence	The Prior
	The probability of getting this evidence if this hypothesis were true	The probability of H being true, before gathering evidence

$$P(H|E) = \frac{P(H|E)\,P(H)}{P(E)}$$

The probability that the hypothesis (H) is true given the evidence (E)

The marginal probability of the evidence (Prob of E over all possibilities)

Imagine, you would like to predict the outcome of a bicycle race between the two cyclists Speedy and Doping. Both adversaries have raced against each other before. Doping won 14 times in the last 20 races.

$$p(Doping) = \frac{14}{20} = 0.7$$

$$p(Speedy) = 1 - 0.7 = \frac{6}{20} = 0.3$$

Based on the information we know Doping has a 30% probability of loosing the race and therefore is more likely the winner of the next race.

Analyzing the previous races in more detail can help to improve the probability calculations even further.

It turns out that Speedy won four of the races when a doping controller show up before the race while Doping only won twice. It seems that Speedy is doing better in “clean” races.

Now, how does this extra piece of information affect our prediction?

We could simple focus on the fact that Doping only won twice in clean races and there is more likely to lose the next race. The probability would of winning the next race would shift from previously 70% to about 33%.

$$p(Doping) = \frac{2}{6} = 0.333$$

This assumption, however, would ignore the fact that Doping has won more competitions in total.

	Race		
	test	no-test	
Doping	2	12	14
Speedy	4	2	6
			20

A better approach would be to combine the two pieces of information and calculate an overall probability for Doping to win the next race.

In probability theory and statistics, the Bayes' theorem describes the probability of an event, based on conditions that might be related to the event. The probability of finding observation A, given that some piece of evidence B is present:

$$p(A|B) = p(B|A)\, p(A) / p(B)$$

For our bicycle race example this would translate into:

$$p(win|test) = p(test|win)\, p(win) / p(test)$$

A - winning the race

B - test before the race

p(A) > p(win) - The probability that Doping wins the next race based on winnings.

p(B) > p(test) - The probability of a doping control before the race.

p(B|A) > p(test|win) - The probability that there is a doping test and Doping wins it.

p(A|B) > p(win|test) - The probability that Doping is winning the race

$$p(win|test) = p(test|win)\ p(win) / p(test)$$

$$p(test|win) = \frac{2}{6} \qquad p(win) = \frac{14}{20} \qquad p(test) = \frac{6}{20}$$

	test	no-test	
win	2	12	14
loose	4	2	6
			20

$$p(win|test) = \frac{\frac{2}{6} \times \frac{14}{20}}{\frac{6}{20}} = \frac{2 \times 14 \times 20}{6 \times 20 \times 6} = \frac{7}{9} = 0.78$$