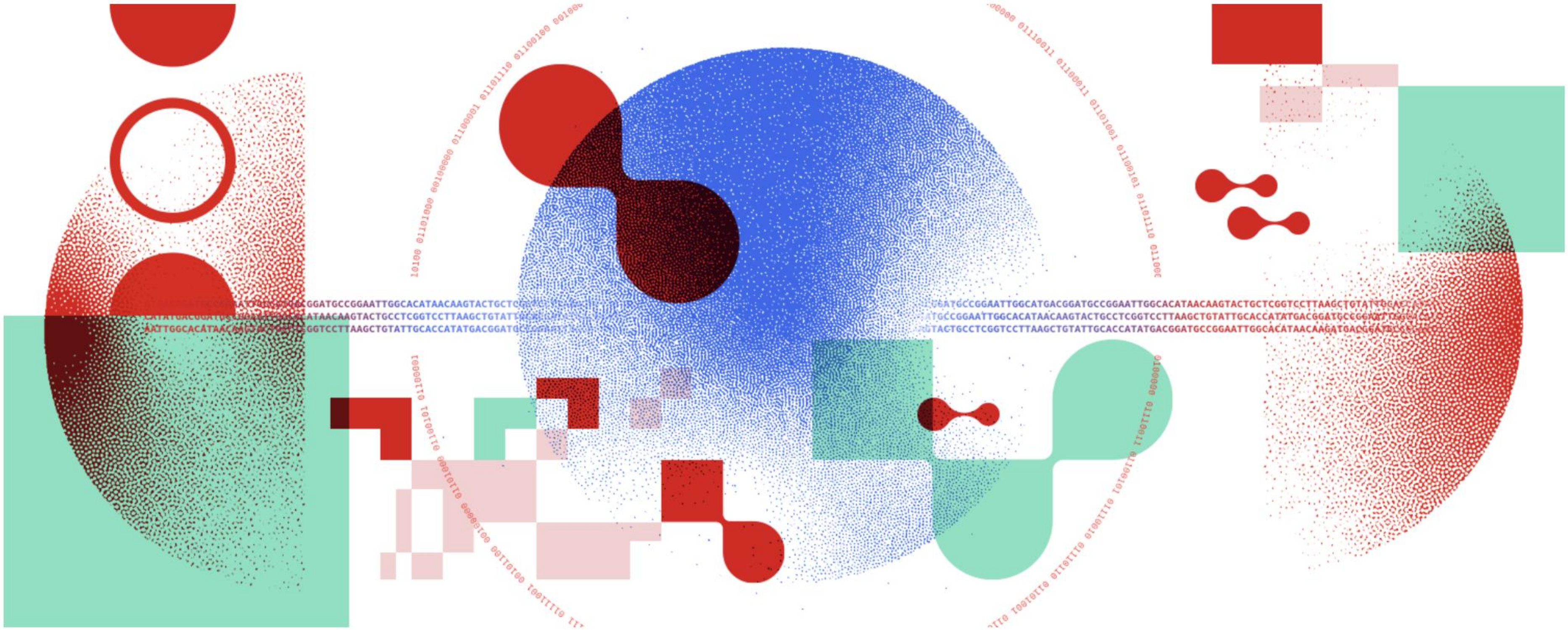


Introduction to Statistics and Data Visualisation with R

Lausanne, January 2026

Joao Lourenço and Rachel Marcone

ANOVA

T-tests: summary

T-test in general
Used to compare means

One-sample t-test
Compare the mean of a sample to a given number

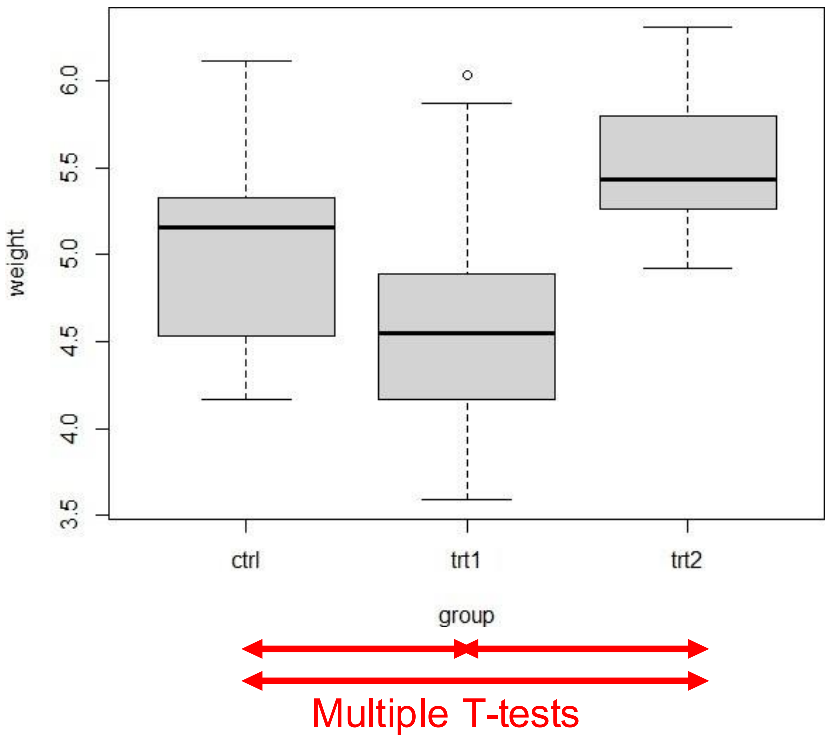
Two-sample t-test
Compare the means of two samples

Paired t-test
Compare the difference between pairs of related data points

One or two groups

How to compare the mean of 3 groups ?

Example: What is the effect of treatment conditions on plant growth (weight) ?



How to compare the mean of 20 groups ?

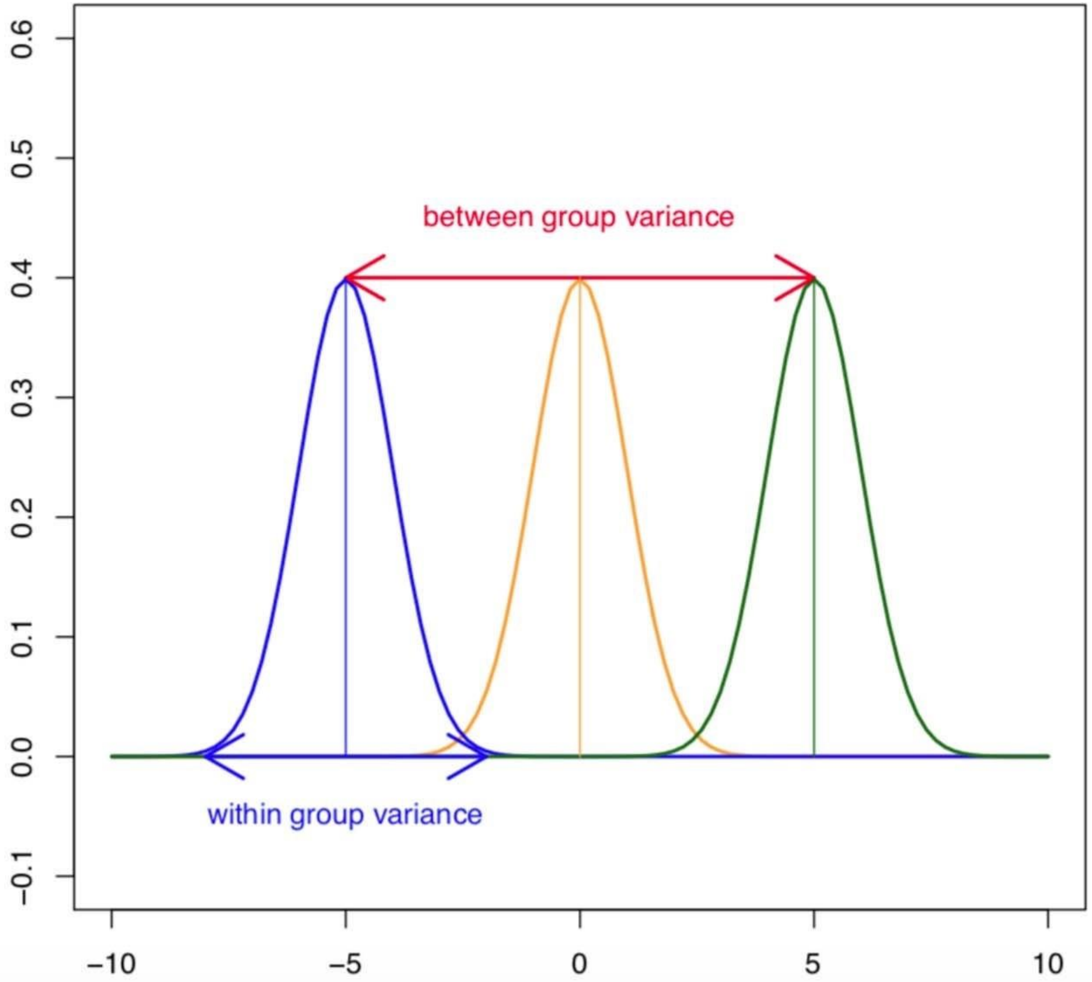
Multiple T-tests → Multiple testing correction !

Another solution ?

ANOVA = ANalysis Of Variance

allows to determine whether there are any statistically significant differences between the means of three or more independent groups

ANOVA – Schematic view



Within group variance = SS_{error}

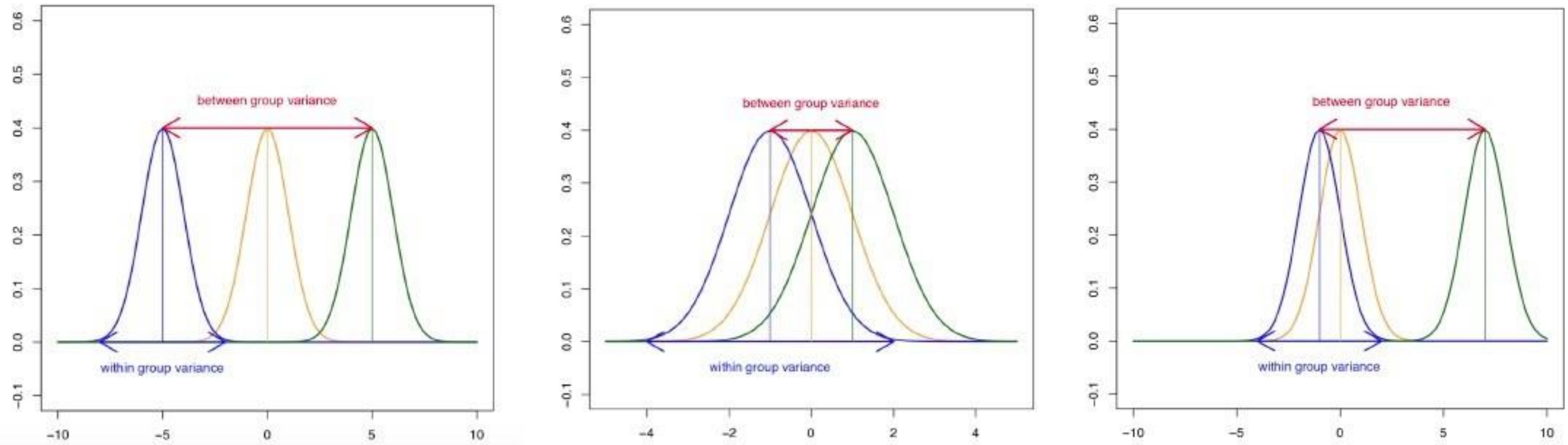
Assumption: SS_{error}=SS_{error}=SS_{error}

Between group variance = SS_{group}

$SS_{\text{total}} = SS_{\text{group}} + SS_{\text{error}}$

ANOVA – Schematic view

If $SS_{\text{group}} > SS_{\text{error}}$ → at least two means are different



ANOVA – Hypothesis testing

- H_0: all group means are equal
- H_1: at least one mean is different

- A simple model formula in R with one factor is written as

```
plant weight ~ treatment
y ~ x
```

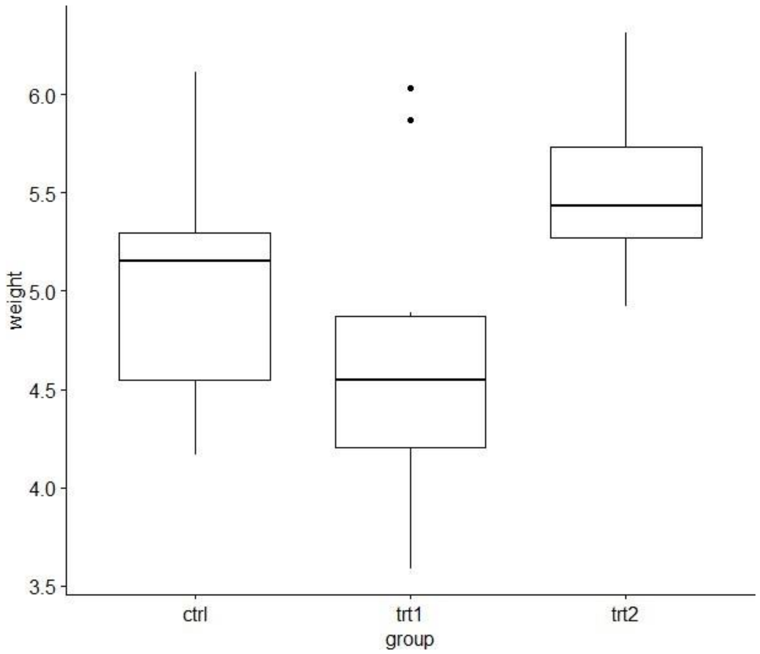
```
modeled by
```

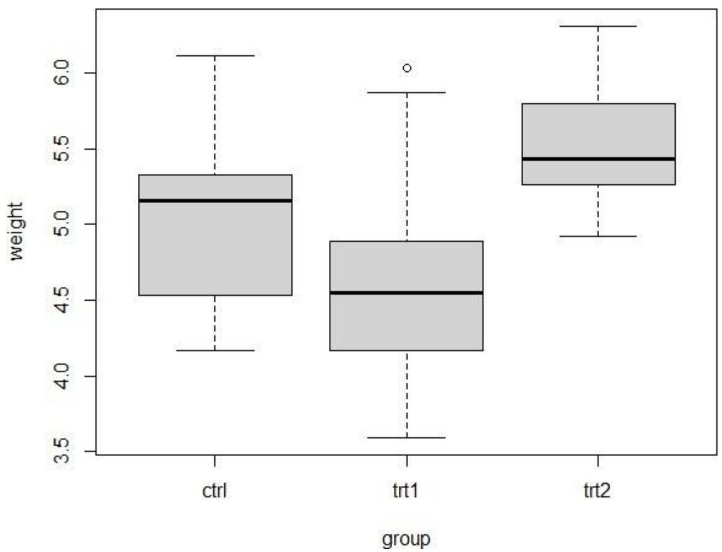
ANOVA – in R

```
# read data
> PlantGrowth <- read.csv("PlantGrowth.csv", header = T)
> dim(PlantGrowth)
> levels(PlantGrowth$group)
> summary(PlantGrowth)

# if the levels are not automatically in the correct order, re-order them as follow:
>PlantGrowth <- PlantGrowth %>% reorder_levels(group, order = c("ctrl", "trt1",
"trt2"))

# compute some summary statistics (count, mean and sd) per group
>PlantGrowth %>% group_by(group) %>% get_summary_stats(weight, type = "mean_sd")
# A tibble: 3 x 5
  group variable       n  mean    sd
  <fct> <chr>      <dbl> <dbl> <dbl>
1 ctrl  weight        10  5.03 0.583
2 trt1  weight        10  4.66 0.794
3 trt2  weight        10  5.53 0.443
```

ANOVA – in R

```
# create a box plot of weight by group:
> ggboxplot(PlantGrowth, x = "group", y = "weight")
> boxplot(PlantGrowth$weight ~ PlantGrowth$group, xlab="group", ylab="weight")
```





ANOVA – in R

```
>anova.res <- aov(PlantGrowth$weight ~ PlantGrowth$group)
Call:
   aov(formula = PlantGrowth$weight ~ PlantGrowth$group)

Terms:
                PlantGrowth$group  Residuals
Sum of Squares            3.76634   10.49209
Deg. of Freedom                 2         27

Residual standard  error: 0.6233746
Estimated effects  may be unbalanced

> summary(anova.res)
                  Df Sum Sq Mean Sq F value Pr(>F)
PlantGrowth$group  2  3.766  1.8832   4.846 0.0159 *
Residuals         27 10.492  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA – in R

```
> summary(anova.res)
                    Df Sum Sq Mean Sq F value Pr(>F)
PlantGrowth$group    2  3.766  1.8832   4.846 0.0159 *
Residuals           27 10.492  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Between groups (factor)	SSB	k-1	MSB=SSB/k-1	F=MSB/MSW
Within groups (error)	SSW	n-k	MSW=SSW/n-k	
Total	SST=SSB+SSW	n-1		

$$SSB = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{\bar{X}})^2 \qquad SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2 \qquad SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{\bar{X}})^2$$

ANOVA assumptions

- Independence of observations
- Equal variance

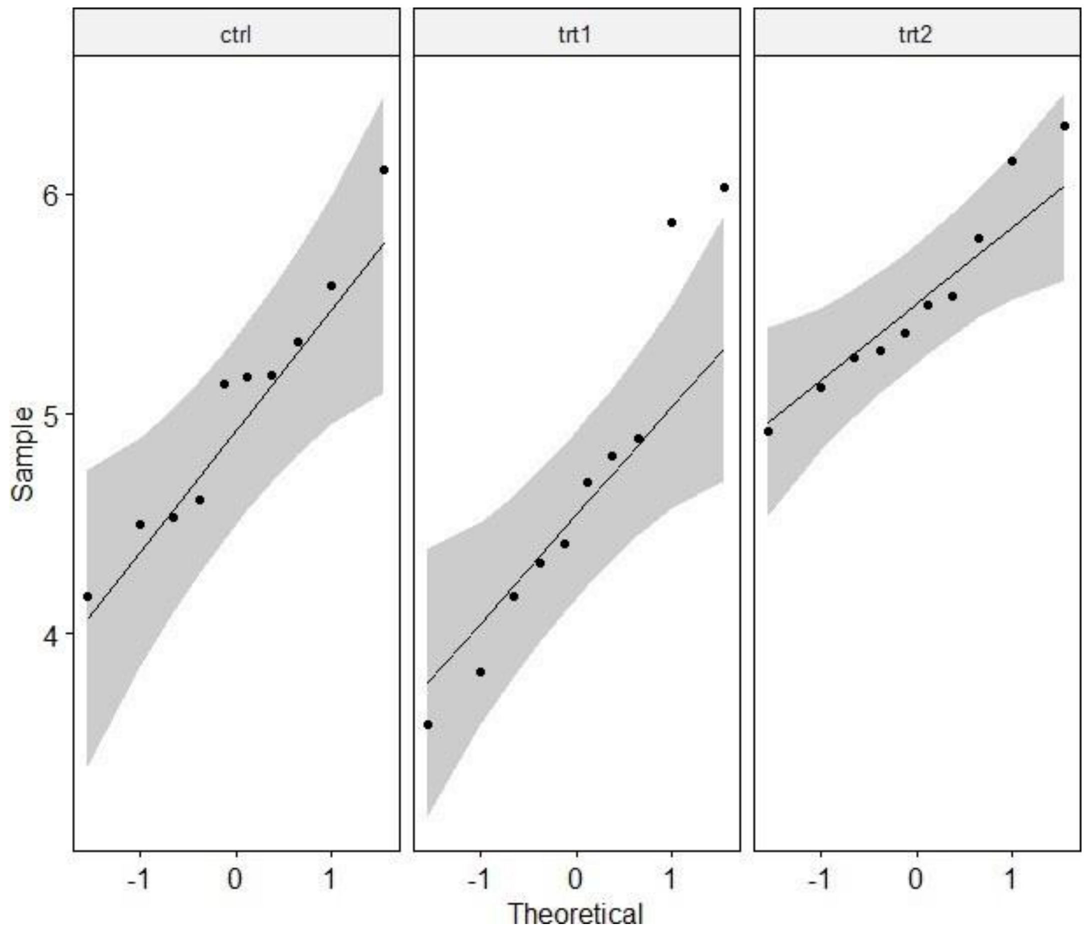
```
>PlantGrowth %>% levene_test(weight ~ group)
# A tibble: 1 x 4
    df1   df2 statistic      p
  <int> <int>     <dbl> <dbl>
1     2    27      1.12 0.341
```

$$W = \frac{n-k}{k-1}\frac{\sum_{i=1}^{k} n_i\left(\frac{1}{n_i}\sum_{j=1}^{n_i}|Y_{ij}-\overline{Y_i}|-\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}|Y_{ij}-\overline{Y_i}|\right)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(|Y_{ij}-\overline{Y_i}|-\frac{1}{n_i}\sum_{j=1}^{n_i}|Y_{ij}-\overline{Y_i}|\right)^2} \sim F(k-1,n-1)$$

ANOVA assumptions

- Normal distribution

```
> PlantGrowth %>% group_by(group)
%>% shapiro_test(weight)
# A tibble: 3 x 4
  group variable  statistic     p
  <fct> <chr>         <dbl> <dbl>
1 ctrl  weight        0.957 0.747
2 trt1  weight        0.930 0.452
3 trt2  weight        0.941 0.564

>ggqqplot(PlantGrowth, "weight",
facet.by = "group")
```



Post-hoc tests

- A significant one-way ANOVA is generally followed up by Tukey post-hoc tests to perform multiple pairwise comparisons between groups

```
>tukey.res <- PlantGrowth %>% tukey_hsd(weight ~ group)
# A tibble: 3 x 9
  term  group1 group2 null.value estimate conf.low conf.high p.adj p.adj.signif
* <chr> <chr>  <chr>       <dbl>    <dbl>    <dbl>     <dbl> <dbl> <chr>
1 group ctrl   trt1            0   -0.371   -1.06      0.320 0.391 ns
2 group ctrl   trt2            0    0.494   -0.197     1.19  0.198 ns
3 group trt1   trt2            0    0.865    0.174     1.56  0.012 *
```

ANOVA is parametric

- ANOVA assumptions
  - Independence of observations
  - Equal variance
  - Normal distribution
- if the above assumptions are not met: non-parametric alternative: Kruskal-Wallis test

```
> kruskal.res <- PlantGrowth %>% kruskal_test(weight ~ group)
> kruskal.res
# A tibble: 1 x 6
  .y.          n  statistic    df        p method
* <chr>    <int>      <dbl> <int>    <dbl> <chr>
1 weight      30       7.99     2  0.0184 Kruskal-Wallis
```

Two-way ANOVA

- Example: the combined effect of treatment type and concentration on the growth (weight) of plants

Treatment type

Concentration

	Control	Treatment 1	Treatment 2
Low			
High			

ANOVA – Hypothesis testing

- A model formula in R with *x* factors is written as

```
y ~ x1+x2+x3
Response ~ predictors
```

- Some useful symbols

`+` add more variables

`-` leave out variables

`:` interaction between two terms

`*` include the terms and the interactions a*b=a+b+a:b

`^n` adds all terms and all interactions up to order n

`I()` include a mathematical expression

Two-way ANOVA

- Example: the combined effect of treatment type and concentration on the growth (weight) of plants

Treatment type

Concentration

	Control	Treatment 1	Treatment 2
Low			
High			

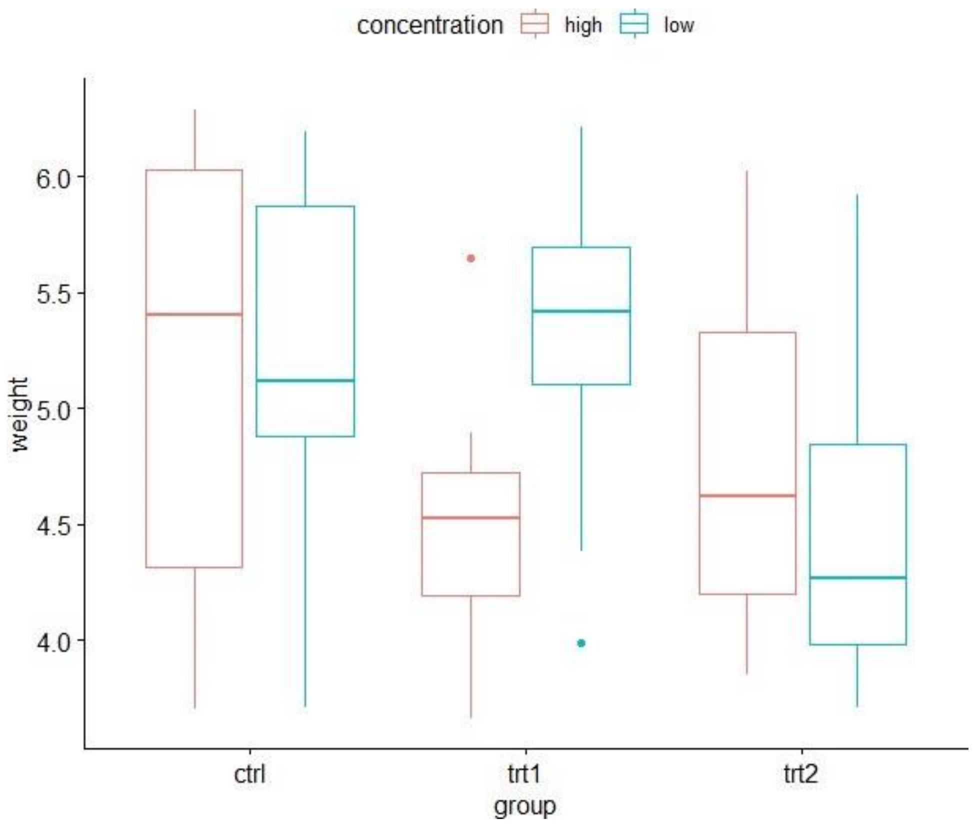
```
Plant growth ~ treatment type * concentration
```

ANOVA – in R

```
# compute some summary statistics (count, mean and sd) per group
>PlantGrowth_new %>% group_by(group, concentration) %>%
get_summary_stats(weight, type = "mean_sd")
# A tibble: 6 x 6
  group concentration  variable      n  mean    sd
  <chr> <chr>          <chr>     <dbl> <dbl> <dbl>
1 ctrl  high           weight       10  5.16 1.00
2 ctrl  low            weight       10  5.24  0.755
3 trt1  high           weight       10  4.51  0.552
4 trt1  low            weight       10  5.30 0.69
5 trt2  high           weight       10  4.77  0.745
6 trt2  low            weight       10  4.55  0.775
```

ANOVA – in R

```
# visualization
> ggboxplot(PlantGrowth_new, x = "group", y = "weight", color = "concentration")
```



ANOVA – in R – check assumptions

- Independence of observations
- Equal variance

```
>PlantGrowth_new %>% levene_test(weight ~ group*concentration)
# A tibble: 1 x 4
    df1   df2 statistic     p
  <int> <int>     <dbl> <dbl>
1     5    54     0.898 0.489
```

ANOVA – in R – check assumptions

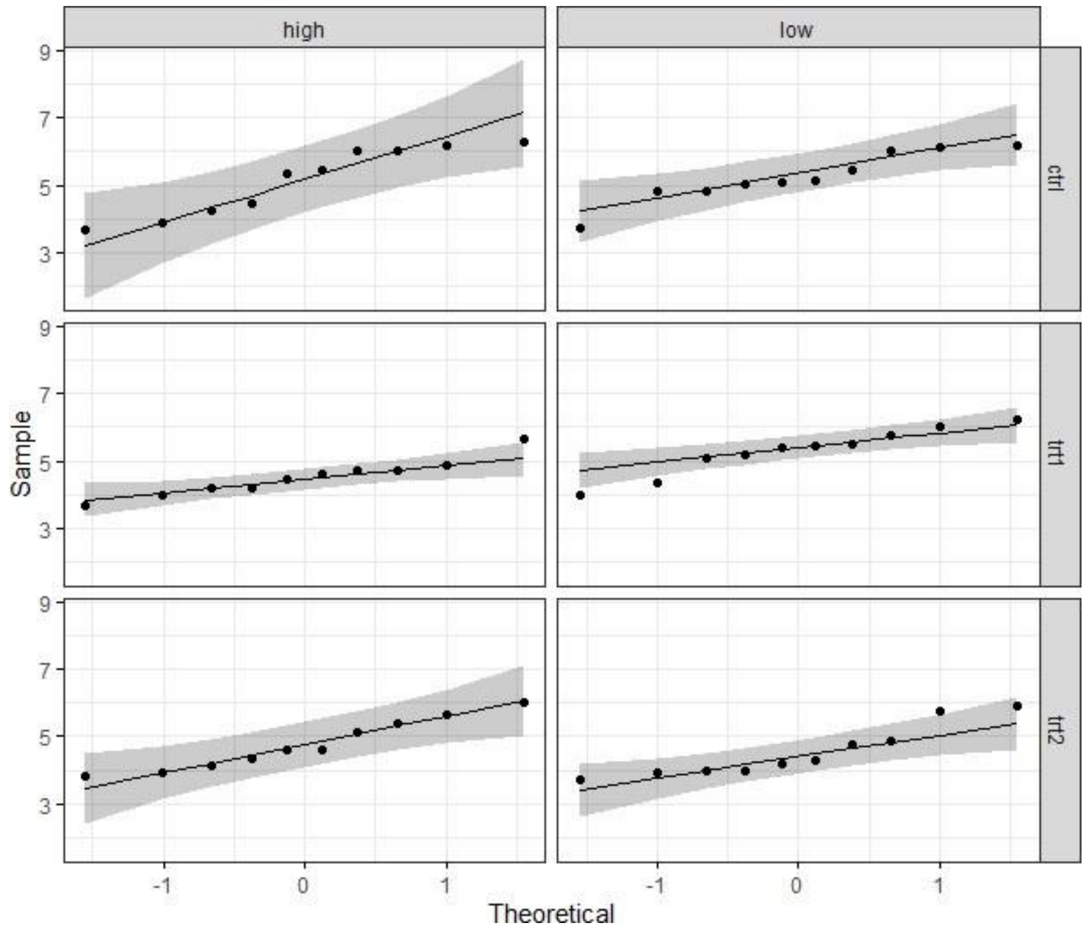
- Normal distribution

```
> PlantGrowth_new %>% group_by(group, concentration) %>% shapiro_test(weight)
# A tibble: 6 x 5
  group concentration variable statistic     p
  <chr> <chr>         <chr>        <dbl> <dbl>
1 ctrl  high          weight       0.883 0.143
2 ctrl  low           weight       0.914 0.313
3 trt1  high          weight       0.963 0.817
4 trt1  low           weight       0.941 0.562
5 trt2  high          weight       0.943 0.585
6 trt2  low           weight       0.867 0.093
```

ANOVA – in R – check assumptions

- Normal distribution

```
>ggqqplot(PlantGrowth_new,
"weight", ggtheme = theme_bw()) +
facet_grid(group ~ concentration)
```



ANOVA – in R

```
>anova.res <- aov(PlantGrowth_new$weight ~ PlantGrowth_new$group *
PlantGrowth_new$concentration)
> summary(anova.res)
                                                     Df Sum Sq Mean Sq F value Pr(>F)
PlantGrowth_new$group                                 2  2.980  1.4898   2.548 0.0876 .
PlantGrowth_new$concentration                         1  0.700  0.6998   1.197 0.2788
PlantGrowth_new$group:PlantGrowth_new$concentration   2  2.734  1.3668   2.338 0.1063
Residuals                                            54 31.575  0.5847
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA – in R

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F ratio
Factor A	SSA	a-1	MSA = SSA/(a-1)	MSA/MSE
Factor B	SSB	b-1	MSB = SSB/(b-1)	MSB/MSE
Interaction	SSAB	(a-1)(b-1)	SSAB = MSAB/(a-1)(b-1)	MSAB/MSE
Error	SSE	ab(n_{ij}-1)	SSE = MSE/(ab(n_{ij}-1))	
Total	SST	n-1		

X_{ijk}: value of kth observation of level i of factor A and level j of factor B
n_i: number of observations of level i of factor A
n_j: number of observations of level j of factor B
n_{ij}: number of observations of level i of factor A and level j of factor B

$$SSA = \sum_{i=1}^{a} n_i(\bar{X}_i - \bar{\bar{X}})^2$$

$$SSB = \sum_{j=1}^{b} n_j(\bar{X}_j - \bar{\bar{X}})^2$$

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}(\overline{X_{ij}} - \bar{X}_i - \bar{X}_j + \bar{\bar{X}})^2$$

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (X_{ijk} - \overline{X_{ij}})^2$$

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{\bar{X}})^2$$

Confidence intervals

Confidence intervals

- Confidence interval is related to the p-value.
- It is a measure of the study's precision.
- P-value answers the question:

"Is there a statistically significant difference between the two treatments ?"

- The point estimate and its confidence interval answer the questions:

"What is the size of that treatment difference?"

"How precisely did this trial determine or estimate the treatment difference?"

Confidence intervals - representation

- Width of a confidence interval:

Confidence Limits: The upper and lower end points of the confidence interval.

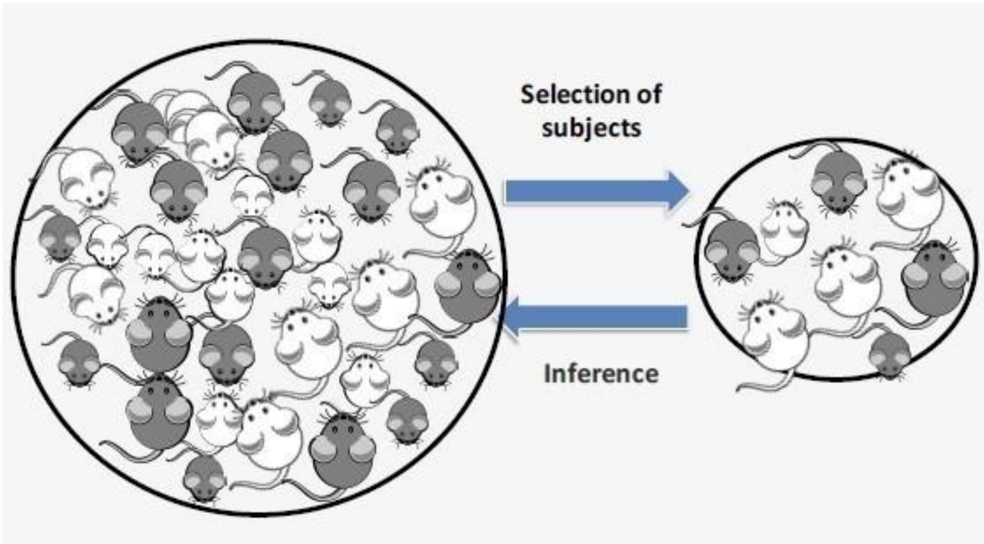
- A narrow CI implies high precision
- A wide CI implies poor precision (usually due to inadequate sample size)

Confidence intervals – computation

- CI = (Sample statistic) ± [(critical value) × (Sampling variability measure)]
  - Sample statistic: observed magnitude of effect or association (e.g., odds ratio, risk ratio, difference in mean)
  - Critical value: reflects on how confident you want to be, related to the statistics and to your level of confidence ($1.0 - \alpha$). The latter is usually expressed as a percentage (e.g. 90%, 95% or 99%). At 95 % the t-statistics critical value is 1.96 for example.
  - Sampling variability: a measure of how high the sampling variability is. Ex: Standard error (S.E.) of the estimate is a measure of variability

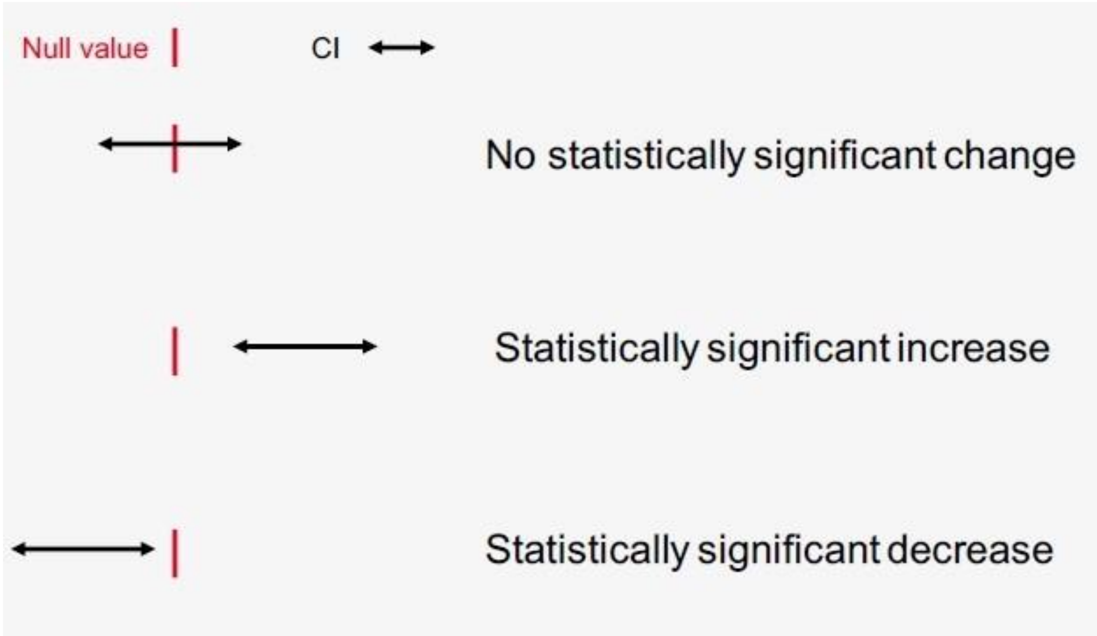
Confidence intervals – interpretation

- 95% C.I. means that true estimate of effect (ex: difference in mean, risk, rate) lies within 1.96 "standard errors" of the population mean 95 times out of 100 (given some assumptions).



Confidence intervals – interpretation

- If the 95% confidence interval does NOT include the null value, then we declare a "statistically significant" association.
- If the 95% confidence interval includes the null value, then the test result is "not statistically significant."



Confidence intervals – interpretation

- Interpretation of C.I. for means: does the interval include 0 ?
- Interpretation of C.I. for ratio: does the interval include 1 ?

- Connection between P-values and C.I.s (they are mathematically connected!)
  - If a 95% CI includes the null effect, the Pvalue is > 0.05 (and we would fail to reject the null hypothesis)
  - If the 95% CI excludes the null effect, the Pvalue is < 0.05 (and we would reject the null hypothesis)

Confidence intervals – interpretation

Exposure:	alcohol intake (high versus low)
Outcome:	Incidence of breast cancer
Risk Ratio:	1.32 (point estimate)
p-value:	0.14 (not statistically significant)
95% C.I.:	0.87 - 1.98

Women with high alcohol intake are 1.32 times (or 32%) more likely to develop breast cancer compared to women with low alcohol intake.
However, we are 95% confident that the true value (risk) of the population lies between
0.87 and 1.98
=> not significant !