17C Laboratory & Professional Skills: Data Analysis

Emma Rand
Data Analysis in R

More than two samples: One-way ANOVA and Kruskal-Wallis

Last week

- Independent and non-independent samples
- Two-sample-tests
  - The two-sample *t*-test
  - The two sample Wilcoxon, also known as the Mann-Whitney
- In RStudio
  - *t*-tests and their non-parametric equivalents
  - Summarising, plotting and reporting

Summary of this week

Extend our ability to test for differences between two or more groups: one-way ANOVA and its non-parametric equivalent Kruskal-Wallis

- Why not do several two-sample tests?
- ANOVA terminology and concepts
- ANOVA assumptions
- Running, interpreting and reporting an ANOVA
- Post-hoc analysis (after a significant ANOVA)
- When assumptions are not met: Kruskal-Wallis
- Running, interpreting and reporting Kruskal-Wallis
- Post-hoc analysis (after a significant Kruskal-Wallis)

Learning objectives for the week

By attending the lectures and practical the successful student will be able to

- Explain the rationale behind ANOVA understand the meaning of the *F* values (MLO 1 and 2)
- Select, appropriately, one-way ANOVA and Kruskal-Wallis (MLO 2)
- Know what functions are used in R to run these tests and how to interpret them(MLO 3 and 4)
- Know how to state the results of these tests scientifically (MLO 3 and 4)
- Create figures for these tests which are suitable for including in a scientific report (MLO 3 and 4)

Review and rationale

Reminder: The choice of test depends on ….

1. Type of data

The type of values a variable can take: Discrete or continuous?

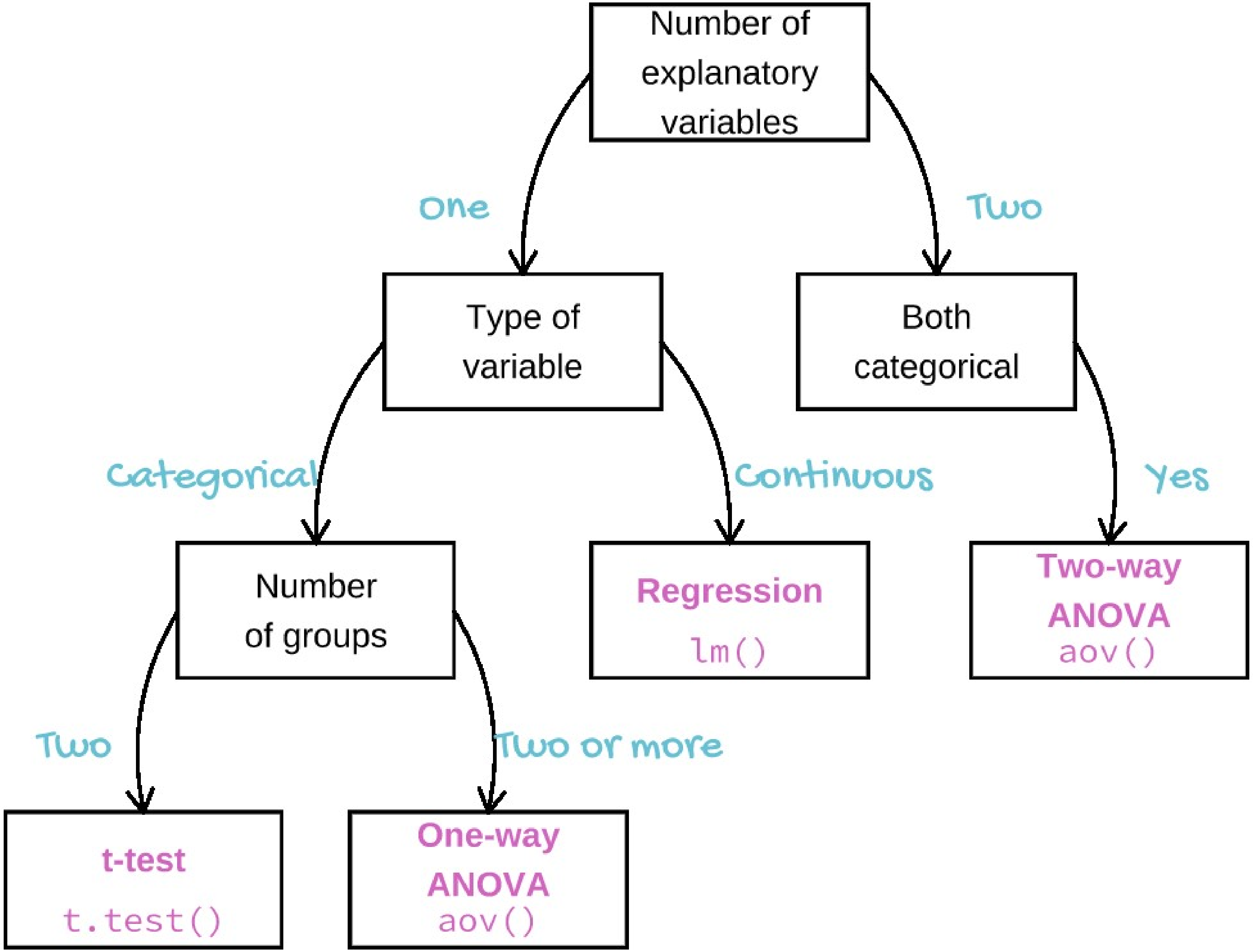
2. Their role in the analysis

Which is the response and which is/are explanatory?

(week 3 Hypothesis testing, data types, reading data in to R and saving figures

Choosing tests: 3 steps

1. What is a one sentence description of what you want to know?
2. What are your explanatory variables?
   - Categories: *t*-tests, ANOVA, Wilcoxon, Mann-Whitney
   - Continuous: Regression, correlation
3. What is your response variable?
   - Normally distributed: *t*-tests, ANOVA, regression
   - Counts: Chi-squared or stage 2 ☺


Number of explanatory variables
One
Two
Type of variable
Both categorical
Categorical
Continuous
Yes
Number of groups
Regression
lm()
Two-way ANOVA
aov()
Two
Two or more
t-test
t.test()
One-way ANOVA
aov()

Why ANOVA, not several *t*–tests?

- Type I error: Rejecting the null hypothesis when it is true
- This will happen with a probability of 0.05
- Doing lots of comparisons increases the type 1 error rate
- ANOVA tests for an effect of the explanatory variable without increasing type 1 error rate

Why ANOVA, not several *t*–tests?

- But, t-tests and ANOVA work in fundamentally the same way
- Both use 'residual' variation to see if explanatory variable (treatment) variation is big

$$t = \frac{statistic - hypothesised\ value}{s.e. of\ statistic}$$

$$F = \frac{Treatment\ MS}{Residual\ MS}$$

One-way ANOVA

Assumptions and alternative

ANOVA, like *t*-tests assumes the “residuals” are normally distributed and have homogeneity of variance

Kruskal-Wallis is the non-parametric equivalent when assumptions are not met.

The one-way ANOVA

A parametric test

One-way ANOVA

Example

- Which growth medium is best for growing bacterial cultures?
- Explanatory variable is type of media: categorical with 3 groups

    Control

    Control + sugar

    Control + sugar + amino acids

– Response variable is colony diameters (mm)

One-way ANOVA

Example

	diameter	medium
1	11.22	control
2	9.35	control
3	9.15	control
4	10.35	control
5	9.63	control
6	10.96	control
7	10.07	control
8	10.40	control
9	10.33	control
10	9.24	control
11	8.90	with sugar
12	10.75	with sugar
13	11.95	with sugar
14	9.85	with sugar
15	10.12	with sugar
16	10.05	with sugar
17	9.60	with sugar
18	10.10	with sugar
19	10.20	with sugar
20	10.88	with sugar
21	10.45	with sugar + amino acids
22	13.19	with sugar + amino acids
23	11.84	with sugar + amino acids
24	13.35	with sugar + amino acids
25	11.22	with sugar + amino acids

One response, one categorical explanatory variable ("one-way ANOVA" or "one-factor ANOVA")

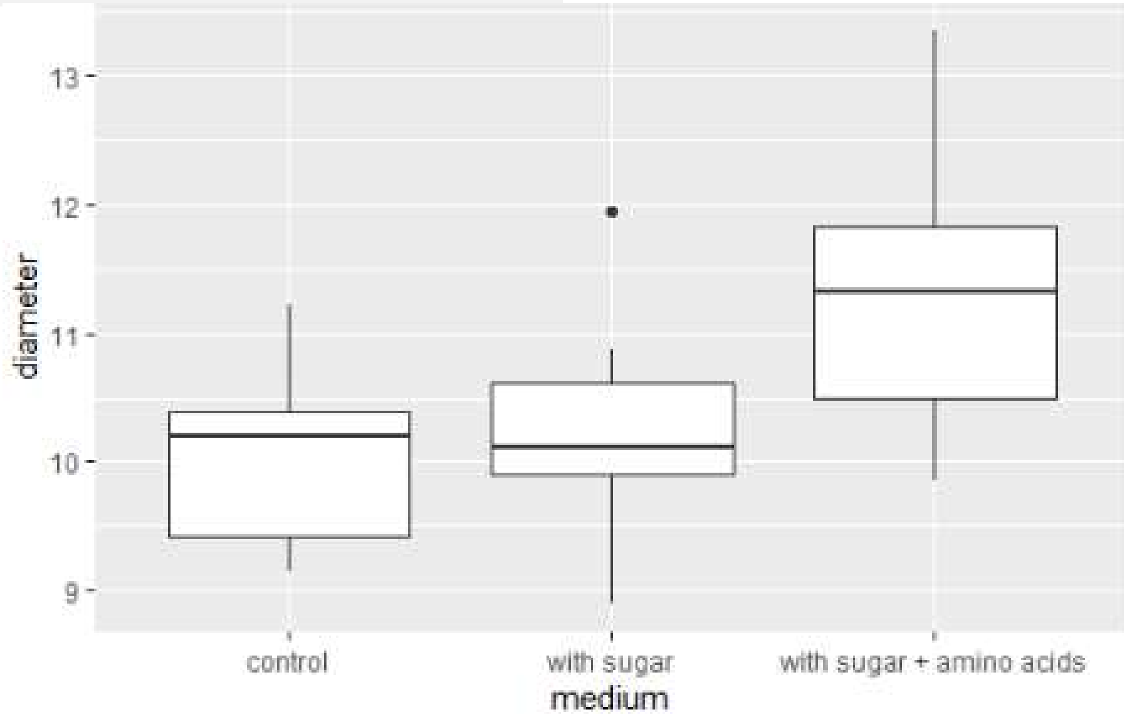
These data are in tidy format.

One-way ANOVA

Example

Plot your data: roughly – perhaps..

```
ggplot(data = culture,
       aes(x = medium, y = diameter)) +
  geom_boxplot()
```



One-way ANOVA

Example

Summarise the data:

```
culturesum <- culture %>%
  group_by(medium) %>%
  summarise(mean = mean(diameter),
            std = sd(diameter),
            n = length(diameter),
            se = std/sqrt(n))
```

```
culturesum
# A tibble: 3 x 5
  medium                     mean   std     n    se
  <fct>                     <dbl> <dbl> <int> <dbl>
1 control                    10.1 0.716    10 0.226
2 with sugar                 10.2 0.818    10 0.259
3 with sugar + amino acids   11.4 1.18     10 0.373
```

One-way ANOVA

Example

Run the anova

```
mod <- aov(data = culture,
           diameter ~ medium)
```

Assign result because we will be able to access residuals from this object later

One-way ANOVA

Example

Name of the dataframe

```
mod <- aov(data = culture,
           diameter ~ medium)
```

The model: explain diameter by medium

One-way ANOVA

Example

Examine the result

P value

```
summary(mod)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
medium        2  10.49   5.247   6.113 0.00646 **
Residuals    27  23.18   0.858
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A key for the line annotation

One-way ANOVA

Terminology

```
              Df Sum Sq Mean Sq F value   Pr(>F)
medium         2  10.49   5.247   6.113 0.00646 **
Residuals     27  23.18   0.858
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sum Sq: “Sums of squares ” (SS): (“*sum squared deviation from the mean*”)

Mean Sq: “Mean square” (MS): variance SS / df (“*average squared deviation from the mean*”)

One-way ANOVA

Terminology

```
              Df Sum Sq Mean Sq F value   Pr(>F)
medium         2  10.49   5.247   6.113 0.00646 **
Residuals     27  23.18   0.858
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Not in output: Total MS: total variation
- 5.247 - Treatment/factor MS: variation due to categorical variable
- 0.858 - Residual MS: background/random/left over variation

One-way ANOVA

Terminology

```
              Df Sum Sq Mean Sq F value  Pr(>F)   
medium         2  10.49   5.247   6.113 0.00646 **
Residuals     27  23.18   0.858                   
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F is the test statistic

It is variable MS / Residual MS

5.247 / 0.858 = 6.113

There is 6.113 times the variance between groups than within them

One-way ANOVA

Checking Assumptions

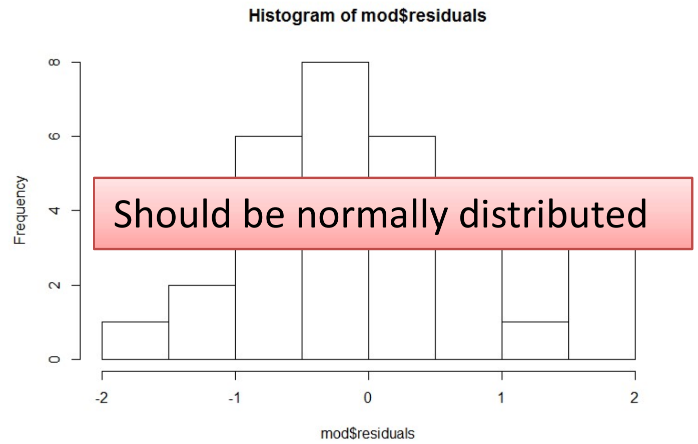
- ANOVA assumes the “residuals” are normally distributed and have homogeneity of variance
- First use common sense: colony diameter is continuous and we would expect it to be normally distributed thus we would expect the residuals to be normally distributed

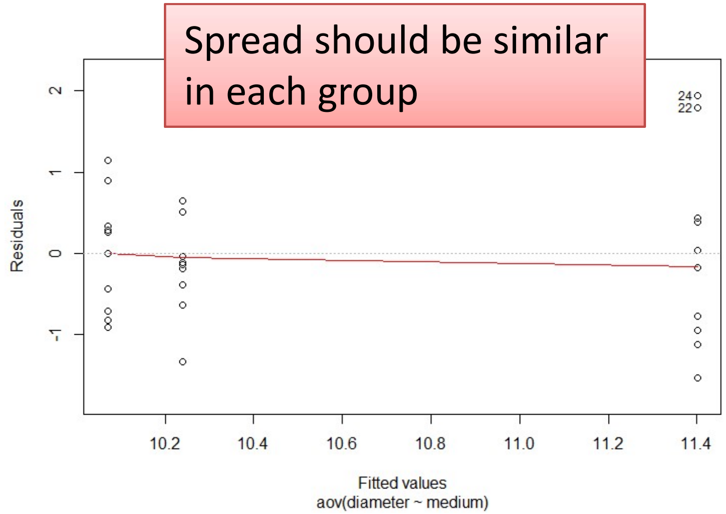
One-way ANOVA

Checking Assumptions

Residuals are calculated for you already!

```
hist(mod$residuals)
shapiro.test(mod$residuals)

  Shapiro-Wilk normality test

data:  mod$residuals
W = 0.96423, p-value = 0.3953

plot(mod, which=1)
```





One-way ANOVA

Example: reporting the result

Reporting the result: "significance, direction, magnitude"

There is a significant effect of media on the diameter of bacterial colonies (ANOVA: F = 6.11; *d.f.* = 2, 27; p = 0.006).

Or

There is a significant difference in diameters between colonies grown on different media (ANOVA: F = 6.11; *d.f.* = 2, 27; P=0.006).

What about direction and magnitude??

Example: direction and magnitude

Which means differ? Post-hoc test needed e.g., Tukey

```
TukeyHSD(mod)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = diameter ~ medium)

$medium
                                      diff       lwr      upr     p adj
with sugar-control                   0.170 -0.857331 1.197331 0.9116894
with sugar + amino acids-control     1.331  0.303669 2.358331 0.0092052
with sugar + amino acids-with sugar 1.161  0.133669 2.188331 0.0243794
```

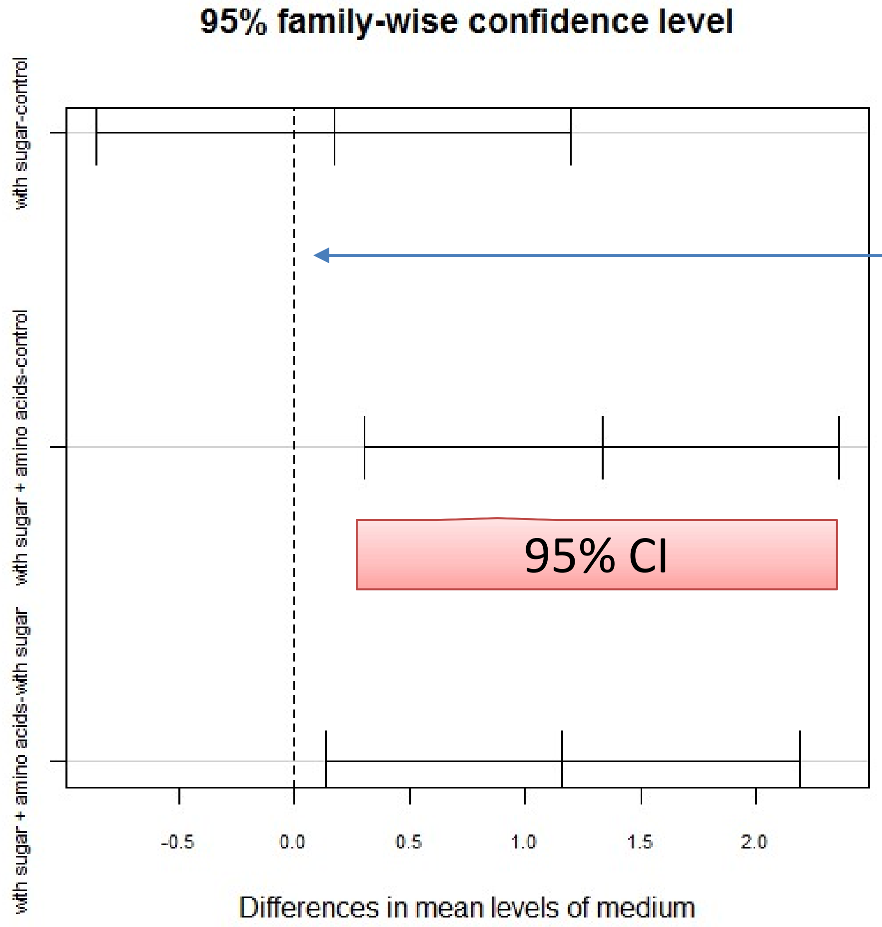
```
                                          diff        lwr      upr     p adj
with sugar-control                       0.170 -0.857331 1.197331 0.9116894
with sugar + amino acids-control         1.331  0.303669 2.358331 0.0092052
with sugar + amino acids-with sugar 1.161  0.133669 2.188331 0.0243794
```

Visualise with post-hoc plot

```
plot(TukeyHSD(mod))
```

A difference of zero

comparison



Example: Reporting the result

There is a significant effect of media on the diameter of bacterial colonies (ANOVA: F = 6.11; $d.f.$ = 2, 27; p = 0.006) with colonies growing significantly better when both sugar and amino acids are added to the medium (see Figure 1).

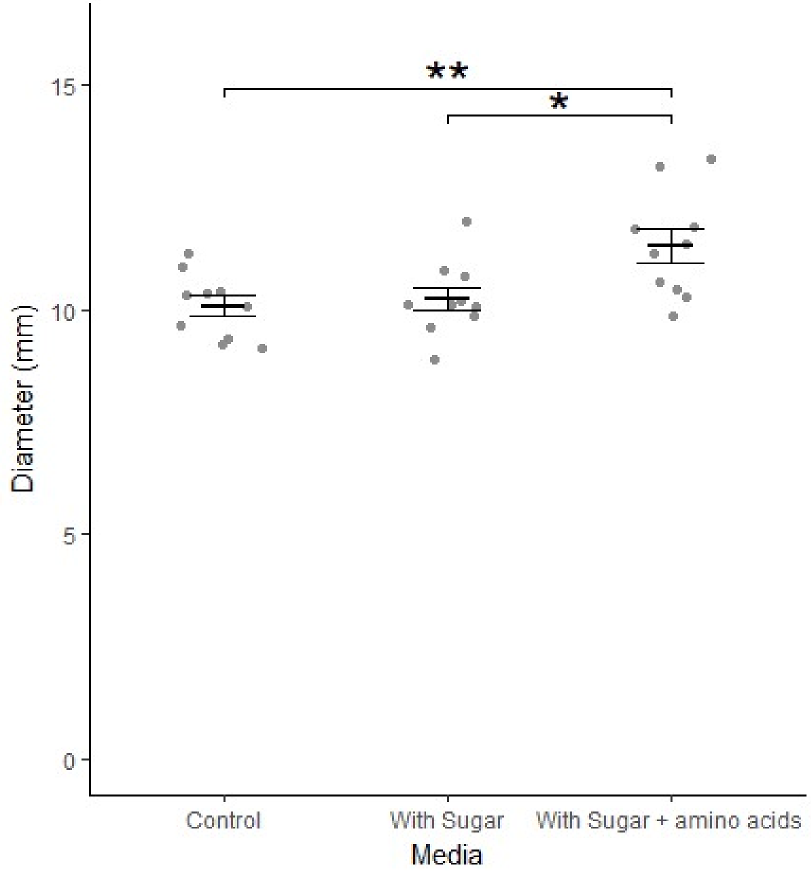


Figure 1. Colony diameter for bacteria grown on different media. Heavy lines are group means with error bars being +/-1 *S.E.* Significant comparisons are indicated.

Example: reporting the result

NOT LIKE THIS!!

There was a significant difference between media and growth rates ……….

It doesn't make sense

One-way ANOVA

Example: reporting the result

There was a significant difference between

$\boxed{\textit{factor levels}}$ in $\boxed{\textit{response}}$ ....

OR.....

There was a significant effect of

$\boxed{\textit{factor}}$ on $\boxed{\textit{response}}$.

One-way ANOVA summary

- Parametric
- To test for a difference between two OR more independent means
- *F* is a variance ration
- Function in R:
  ```
  mod <- aov(data = df, response ~ explanatory)
  summary(mod)
  ```
- If $p < 0.05$ the test is significant
- assumptions: normally and homogenously distributed residuals

continued

One-way ANOVA summary

- ANOVA tells at lest two means differ and a post-hoc test is need to determine which means differ
- Tukey Honest Significant Difference is the post-hoc we used
- Function in R:
  `TukeyHSD(mod)`
- Significance, direction, magnitude
- Figure: data and ‘model’

Kruskal-Wallis

Non-parametric equivalent of the one-way ANOVA

One-way ANOVA

Non-parametric equivalent: Kruskal Wallis

When assumptions are not met

- Residuals not normal
- Unequal variance

Likely when:

- Repeated values
- Small sample size
- Unequal sample size

Non-parametric equivalent of one-way ANOVA

Kruskal Wallis: example on same data

- Same data – to compare power
- Test statistic follows a chi-squared distribution

```
kruskal.test(data = culture, diameter ~ medium)
        Kruskal-Wallis rank sum test

data:  diameter by medium
Kruskal-Wallis chi-squared = 8.1005, df = 2, p-value = 0.01742
```

There is a significant effect of media on diameter

Non-parametric equivalent of one-way ANOVA

Kruskal Wallis: example on same data

Which groups differ? Post-hoc test needed e.g., kruskalmc() in pgirmess package

True = significant

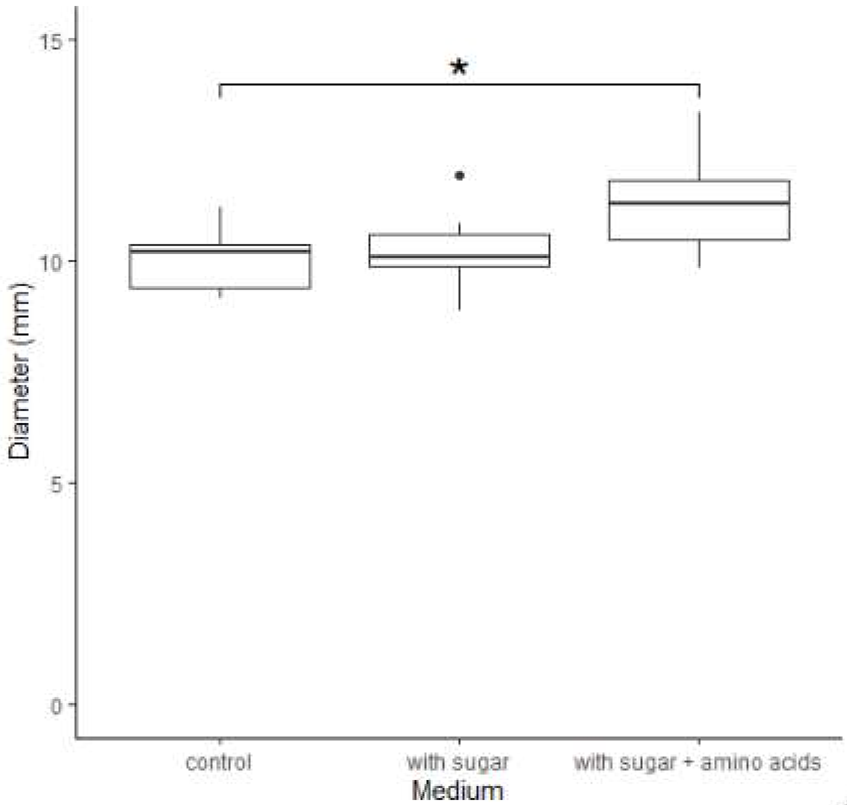
```
library(pgirmess)
kruskalmc(data = culture, diameter ~ medium)

Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
                                         obs.dif critical.dif difference
control-with sugar                          0.85     9.425108      FALSE
control-with sugar + amino acids           10.10     9.425108       TRUE
with sugar-with sugar + amino acids         9.25     9.425108      FALSE
```

Non-parametric equivalent of one-way ANOVA

Kruskal Wallis: example on same data

Reporting the result: “significance, direction, magnitude”

There is a significant effect of media on the diameter of bacterial colonies (Kruskal-Wallis: χ^2= 8.1; *d.f.* = 2; *p* =0.017) with a significant difference only between the control and when sugar and amino acids are added to the medium (see Figure 1).



Kruskal-Wallis summary

- Non-parametric
- when assumptions for one-way ANOVA not met
- To test whether the mean ranks differ
- Function in R:
  `kruskal.test(data = df, response ~ explanatory)`
- If $p < 0.05$ the test is significant
- Few assumptions
- Figure: boxplot