Laboratory & Professional Skills:
Data Analysis

Laboratory & Professional skills for Bioscientists
Term 2: Data Analysis in R

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

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 and complete a partially filled ANOVA table (MLO 1 and 2)
- Apply (appropriately), interpret and evaluate the legitimacy of, one-way ANOVA and Kruskal-Wallis including post-hoc tests in R (MLO 2, 3 and 4)
- Summarise and illustrate with appropriate R figures test results scientifically (MLO 3 and 4)

Choosing tests

filecopy.R* × Untitled1* ×

Filter

	mass	sex
1	18.3	females
2	22.1	females
3	22.4	females
4	18.5	females
5	22.2	females
6	19.3	females
7	17.8	females
8	20.2	females
9	22.1	females
10	16.6	females
11	20.7	females
12	18.7	females
13	22.6	females
14	21.5	females
15	21.7	females
16	19.9	females
17	23.1	females
18	17.8	females
19	19.5	females
20	24.6	females
21	22.7	males
22	20.6	males
23	25.4	males

Showing 1 to 23 of 40 entries

Two groups: two-sample *t*-test

Three groups: ANOVA

Untitled1* × Untitled2* ×

Filter

	values	population
1	10.31	A
2	13.07	A
3	10.33	A
4	10.52	A
5	11.67	A
6	7.27	A
7	10.31	B
8	13.07	B
9	10.33	B
10	10.52	B
11	11.67	B
12	7.27	B
13	10.31	C
14	13.07	C
15	10.33	C
16	10.52	C
17	11.67	C
18	7.27	C

But why not just do 3 2-sample *t*-tests? Type I errors

Why ANOVA, not several *t*–tests?

- Type I error: Rejecting the null hypothesis when it is true (revision lecture 2)

  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

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")

These data are in tidy format:

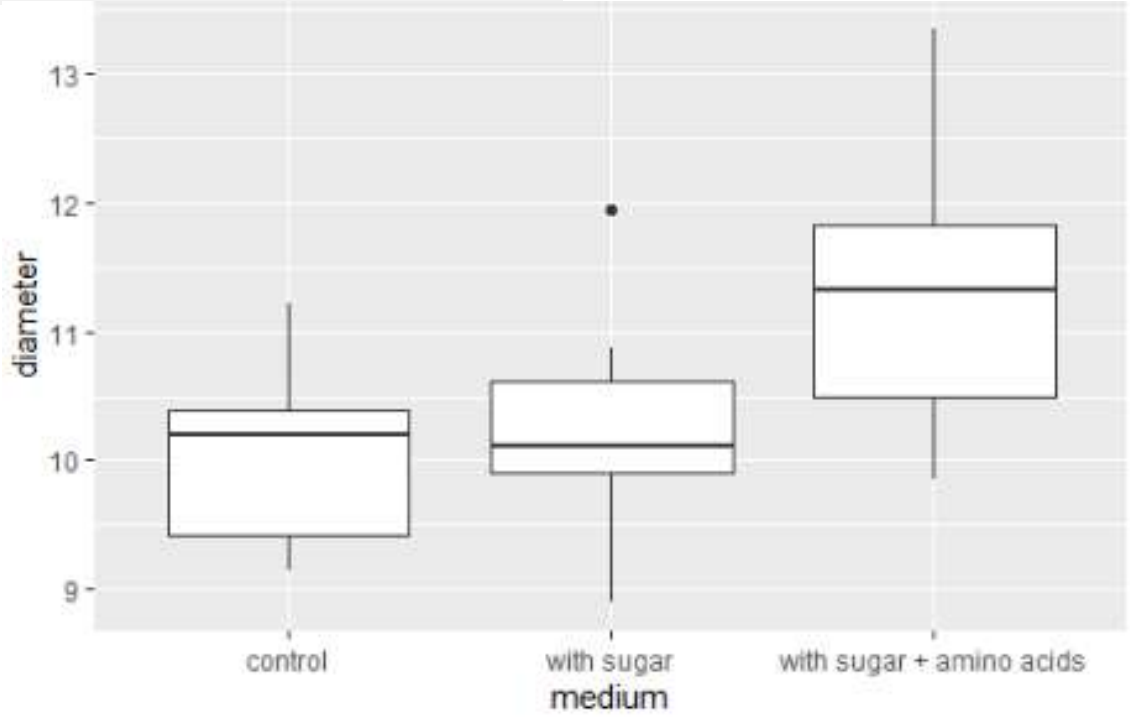
One response per row (all responses in the same column)

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)
```

Name of the dataframe

The model: explain diameter by medium

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

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*”)

See lecture 4

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 factor MS / Residual MS

5.247 / 0.858 = 6.113

There is 6.113 times the variance between groups than within them

Checking Assumptions

Also in previous lecture

- Common sense
  - response should be continuous
  - No/few repeats
- Plot the residuals
- Using a test in R

One-way ANOVA

Checking Assumptions

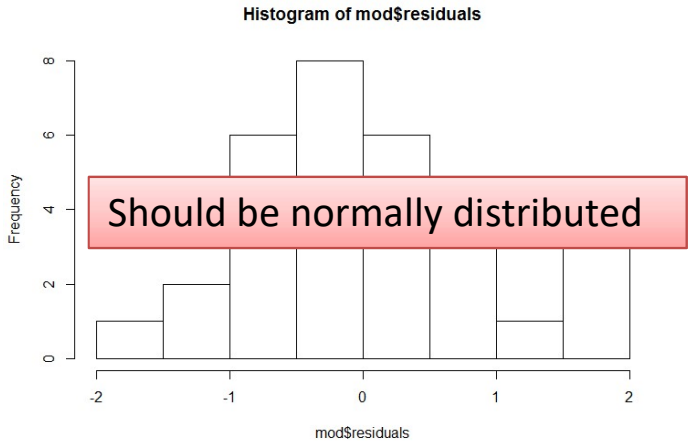
Residuals are calculated for you already!

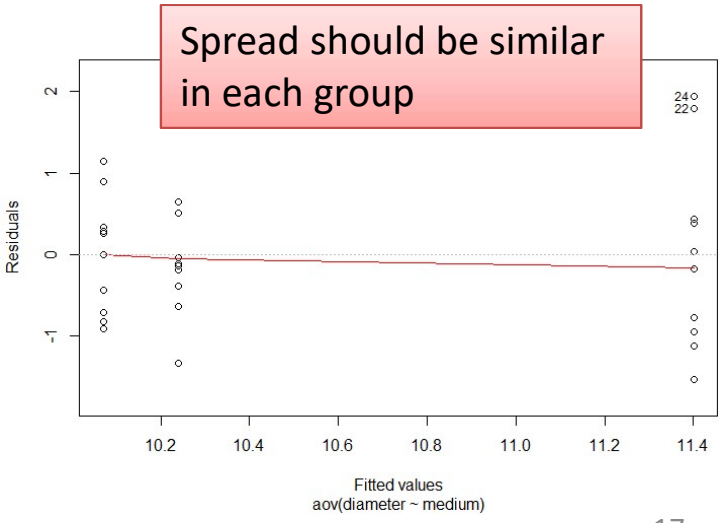
```
hist(mod$residuals)
shapiro.test(mod$residuals)

   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??

One-way ANOVA

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
```

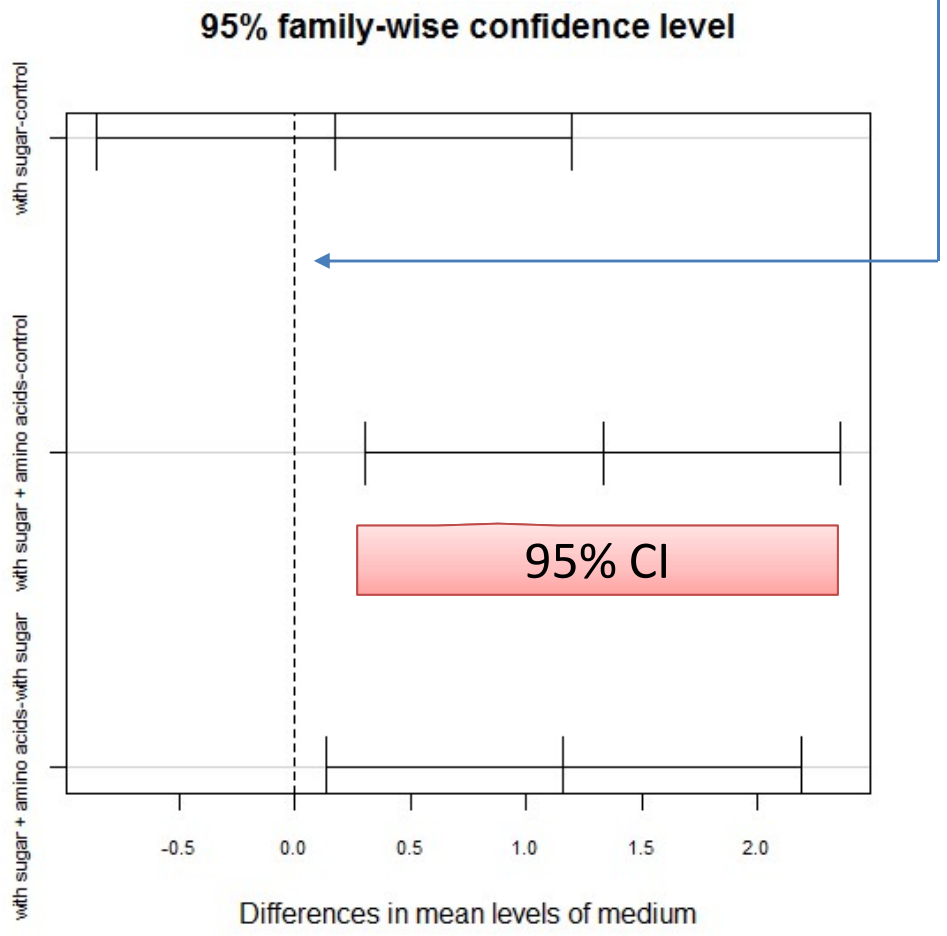
```
                                        diff       lwr      upr     p adj
with sugar-control                     0.170 -0.857331 1.197331 0.9116894
with sugar + amino acids-control       1.331  0.303669 2.358331 0.0092052
with sugar + amino acids-with sugar    1.161  0.133669 2.188331 0.0243794
```

Visualise with post-hoc plot

A difference of zero

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

comparison



One-way ANOVA

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).

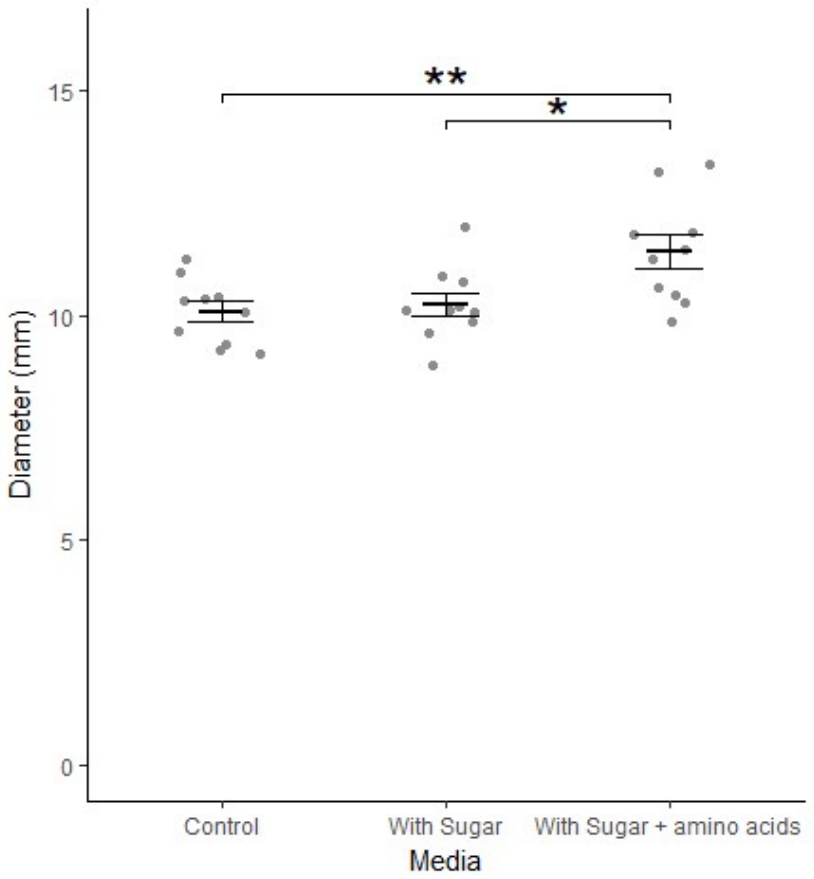


Figure 1. Colony diameter for bacteria grown on different media. Heavy lines are group means with error bars being +/-1 *S.E.* Significant comparisons are indicated.

Example: reporting the result

NOT LIKE THIS!!

There was a significant difference between

media and growth rates …….

It doesn’t make sense

Example: reporting the result

There was a significant difference between

factor levels in *response* ……

OR…..

There was a significant effect of

factor on *response* .

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

```
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
```

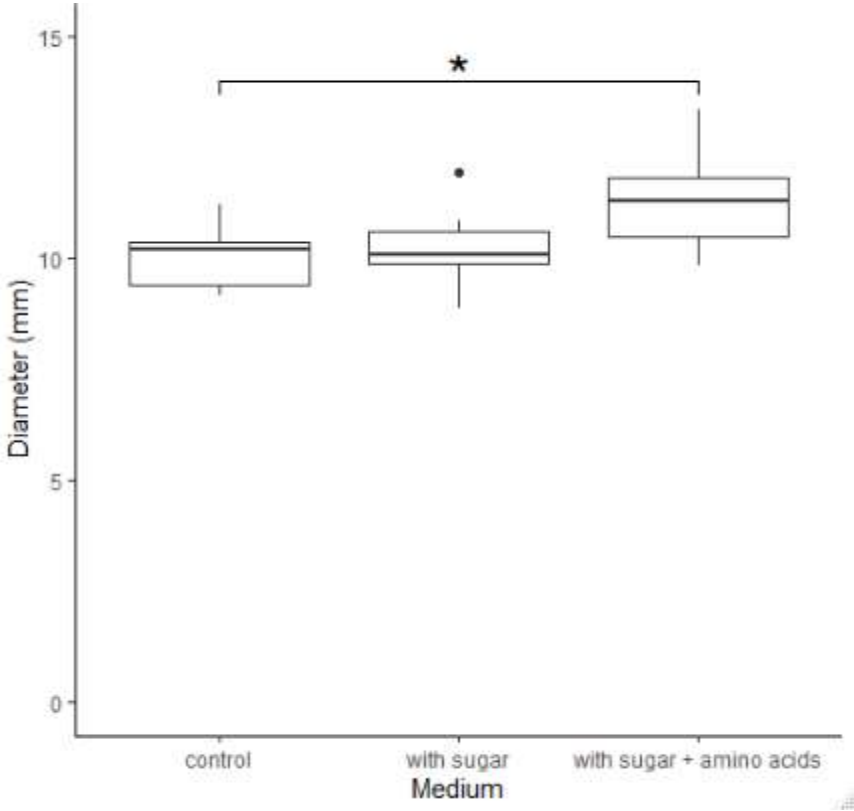
True = significant

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).



Learning objectives for the week

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

- Explain the rationale behind ANOVA and complete a partially filled ANOVA table (MLO 1 and 2)
- Apply (appropriately), interpret and evaluate the legitimacy of, one-way ANOVA and Kruskal-Wallis including post-hoc tests in R (MLO 2, 3 and 4)
- Summarise and illustrate with appropriate R figures test results scientifically (MLO 3 and 4)