

Topic 4: Assumptions, Multiple Comparisons


ENVX2001 Applied Statistical Methods

Aaron Greenville



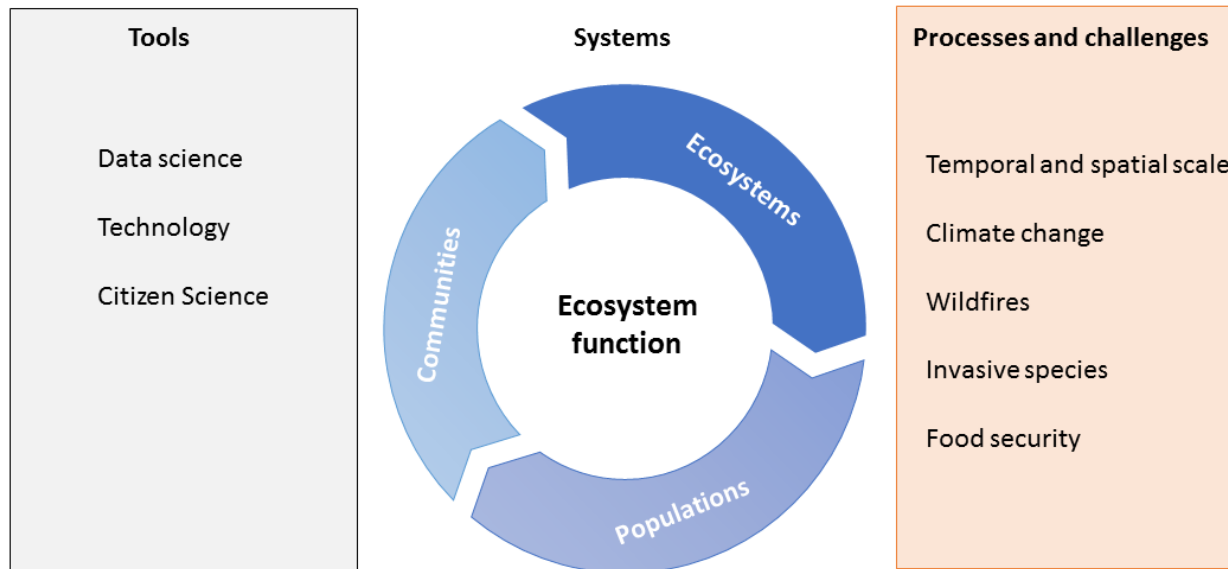
THE UNIVERSITY OF
SYDNEY





We acknowledge the tradition of custodianship and law of the Country on which the University of Sydney campuses stand. We pay our respects to those who have cared and continue to care for Country.

About me



www.aarongreenville.com



Introduction

Learning Outcomes

- Demonstrate proficiency in designing sample schemes and analysing data from them using R;
- Describe and identify the basic features of an experimental design; replicate, treatment structure and blocking structure;
- **Demonstrate proficiency in the use of the statistical programming language R to apply an ANOVA** and fit regression models to experimental data;
- Demonstrate proficiency in the use of the statistical programming language R to use multivariate methods to find patterns in data
- **Interpret the output and understand conceptually how it is derived of a regression, ANOVA and multivariate analysis that have been calculated by R;**
- Write statistical and modelling results as part of a scientific report;
- Appraise the validity of statistical analyses used in publications.

Introduction

Outline

- Part 1: Assumptions
- Part 2: Identifying significant differences between pairs of groups (treatments)
- Part 3: Example walkthrough

Introduction

Topic 4 Learning Outcomes

- At the end of this topic students should be able to:
 - Assess whether a fitted ANOVA model meets its statistical assumptions;
 - Identify which pair(s) of treatment means are significantly different;
 - Demonstrate proficiency in the use of R (and interpretation of the output) for performing the above tasks.

Experimental and statistical modelling design process

- From:
- Fox, G. A., S. Negrete-Yankelevich, and V. J. Sosa. (2015). *Ecological statistics: contemporary theory and application*. Oxford University Press, USA.

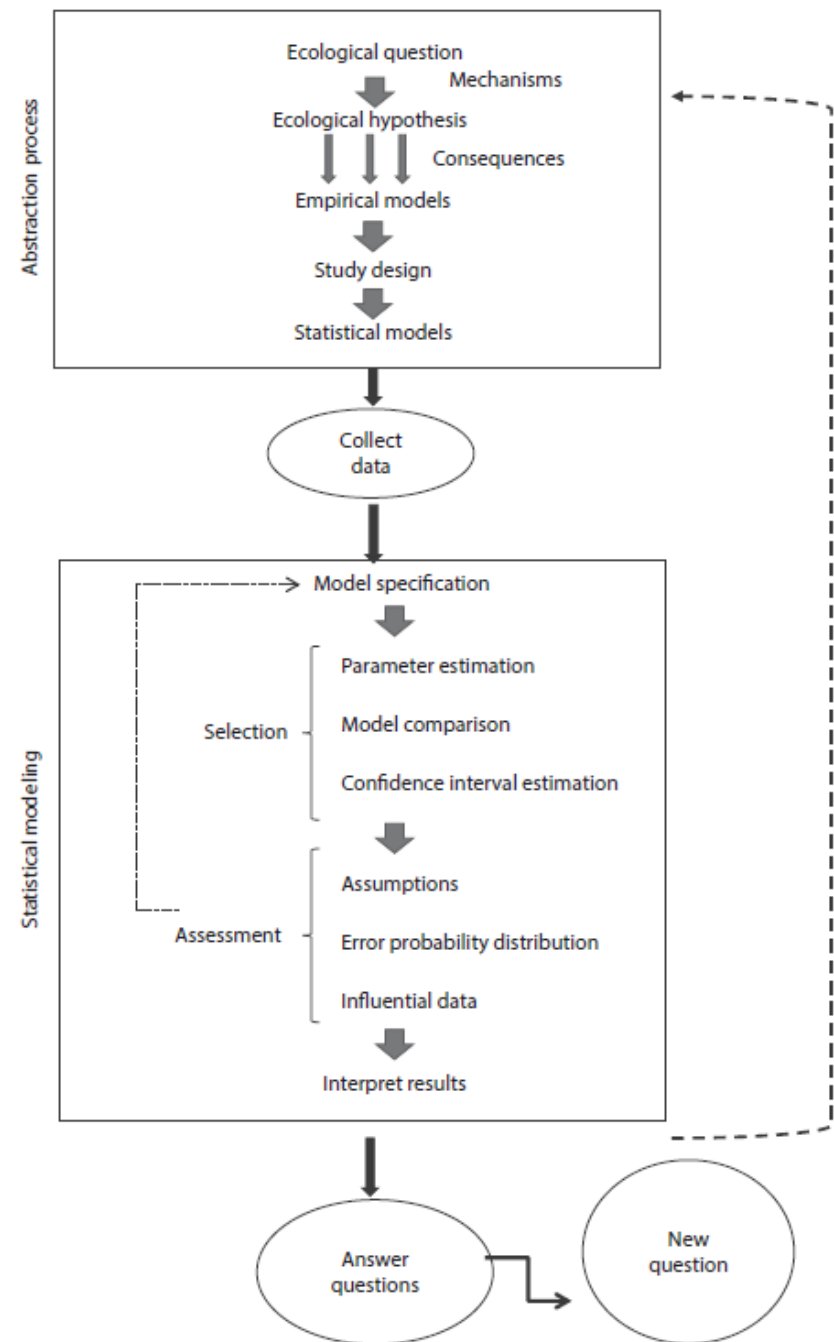


Fig. I.1 The cycle of ecological research and the role of statistical modeling.

Assumptions

Model Equation

- $y_{i,j} = \mu_i + \epsilon_{i,j}$
- Observed data = Group Mean + **Random Error**
- Two assumptions:
 - Random errors are assumed to be normally distributed
 - Random errors assume an equal variance,
 - $N(\mu_i, \sigma^2)$ or $\epsilon_{i,j} \sim N(\mu_i, \sigma^2)$
- For one-way ANOVAs, assumptions can be checked:
 - 1. either directly from the raw data, or
 - 2. from the **residuals**.

Assumptions

Residuals

- Residuals is an estimate of the random error
 - since $y_{i,j} = \mu_i + \epsilon_{i,j}$
 - $\epsilon_{i,j} = y_{i,j} - \mu_i$
 - i.e. residual = observation – fit (prediction)
- \bar{y}_i (group mean) is the **fitted value** in that it is the best estimate of $y_{i,j}$

Assumptions

Chicks example

summary(model.aov)

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Diet	3	16467	5489	6.647	0.004 **
Residuals	16	13212	825.8		

Data	Fit	Residual
Diet 1		
99	79.0	20.0
88	79.0	9.0
76	79.0	-3.0
38	79.0	-41.0
94	79.0	15.0
Mean =		0
SD =		24.5

Diet 3		
42	81.4	-39.4
97	81.4	15.6
81	81.4	-0.4
95	81.4	13.6
92	81.4	10.6
Mean =		0
SD =		22.9

Data	Fit	Residual
Diet 2		
61	71.0	-10.0
112	71.0	41.0
30	71.0	-41.0
89	71.0	18.0
63	71.0	-8.0
Mean =		0
SD =		31.0

Diet 4		
169	142.8	26.2
137	142.8	-5.8
169	142.8	26.2
85	142.8	-57.8
154	142.8	11.2
Mean =		0
SD =		34.9

Note: Mean residual for each Diet is zero, SDs same as for raw data.

Assumptions

```
summary(model.aov)
```

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)	
Diet	3	16467	5489	6.647	0.004	**
Residuals	16	13212	825.8			

Interpretation of Residuals

- For a normal distribution, $\sim 95\%$ of values are within $\pm 1.96 \times \text{SD}$ of mean
- For current model, $s = \sqrt{\text{Residual MS}} = \sqrt{825.8} = 28.7\text{g}$
- So if $y_{ij} \sim N(\mu_i, \sigma^2)$ or $\epsilon_{i,j} \sim N(\mu_i, \sigma^2)$
 - Expect 95% residuals in range $\pm 1.96 \times 28.7 = \pm 56.3\text{g}$
- Only 1 residual exceeds, flagged as potential outlier
 - Diet 4: Observation = 85g, Fit = 142.8g, Residual = -57.8g
- This is just outside range – and one moderate outlier in 20 observations would be expected > > no actions needs to be taken

Assumptions

Standardised Residuals

- To automatically assess if residual is extreme, calculate standardise residuals
- Standardised residuals will be approximately $N(0,1)$ if no extreme outliers
 - 95% of standardised residuals in range ± 1.96 , or say ± 2
- Standardised residual = $\frac{\text{Residual}}{\text{Std.Error residual}}$ where
- Std. Error residual = $\sqrt{\text{Residual MS} \times \frac{n_i - 1}{n_i}}$
 - $\sqrt{\frac{n_i - 1}{n_i}}$ is an adjustment for small sample size

Assumptions

Standardised Residuals

- Example; $Residual\ MS = 825.8, n_4 = 5$
- Std. Error residual = $\sqrt{Residual\ MS \times \frac{n_i - 1}{n_i}} = \sqrt{825.8 \times \frac{5 - 1}{5}} = 25.67\ g$
- Standardised residual = $\frac{Residual}{Std.Error\ residual} = \frac{-57.8}{25.67} = -2.25$ (absolute val. > 2)
- R functions
 - `fitted` returns fitted values from `aov` object
 - `resid` returns residuals from `aov` object
 - `rstandard` returns standardised residuals from `aov` object

Assumptions

› Normality

- Check: Histograms, QQPlots, Normality Test

› Equal variances

- SD/Boxplot of residuals for **each group**
- plot of residuals versus fitted values

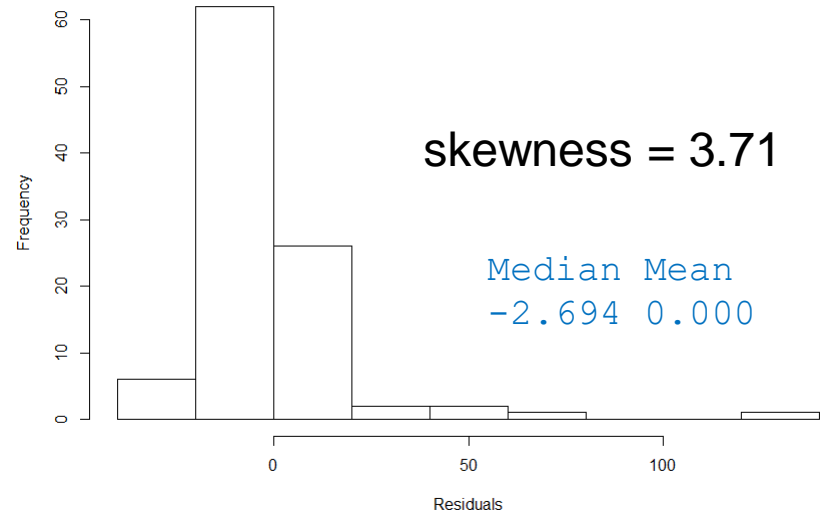
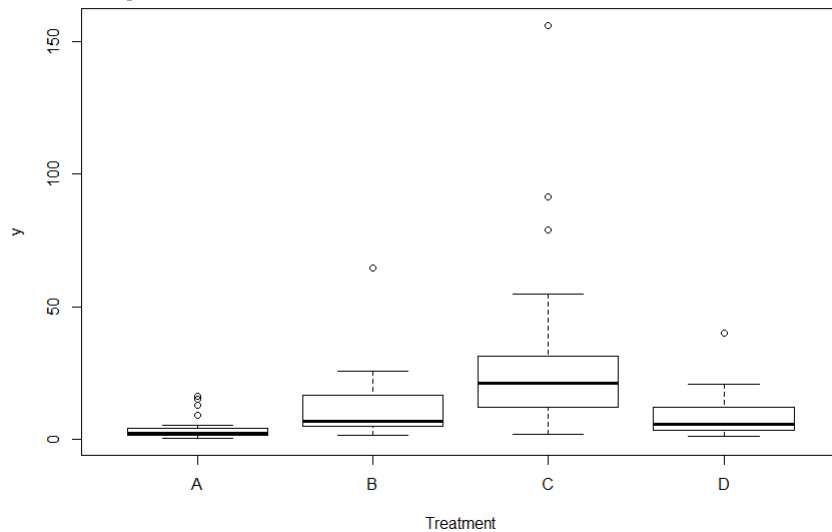
› Independent samples

- Experimental design

Assumptions

Normality

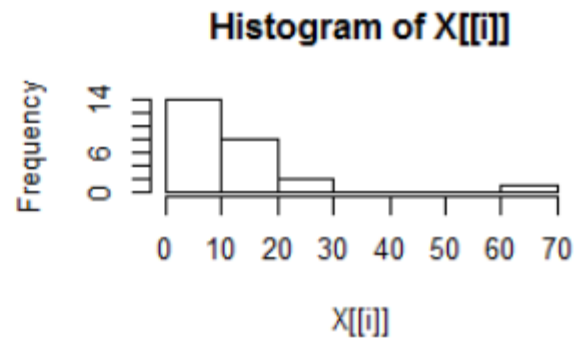
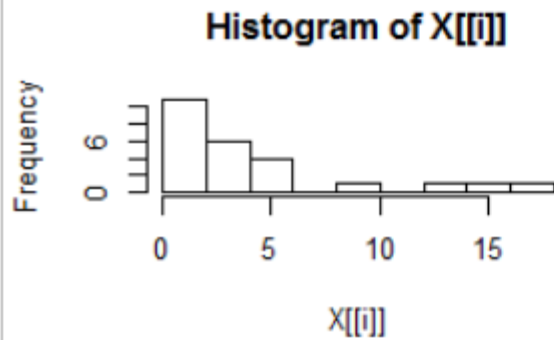
- Synthetic Data - 1 treatment factor, 4 treatment levels (A, B, C, D), 25 replicates



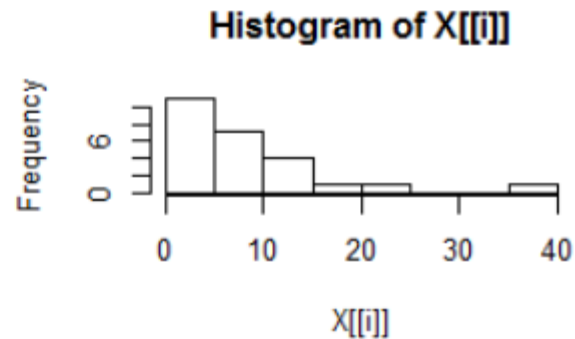
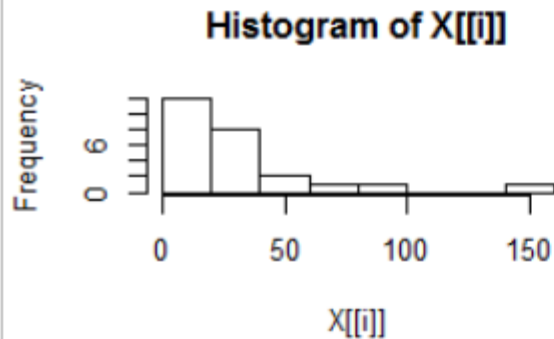
```
> tapply(y, Treatment, median)
      A      B      C      D 
2.185894 6.945461 21.178875 5.690122
```

```
> tapply(y, Treatment, mean)
      A      B      C      D 
4.113848 12.234283 31.387065 8.418323
```

Assumptions



```
> par(mfrow = c(2,2))  
> tapply(y,Treatment,hist)  
> par(mfrow = c(1,1))
```

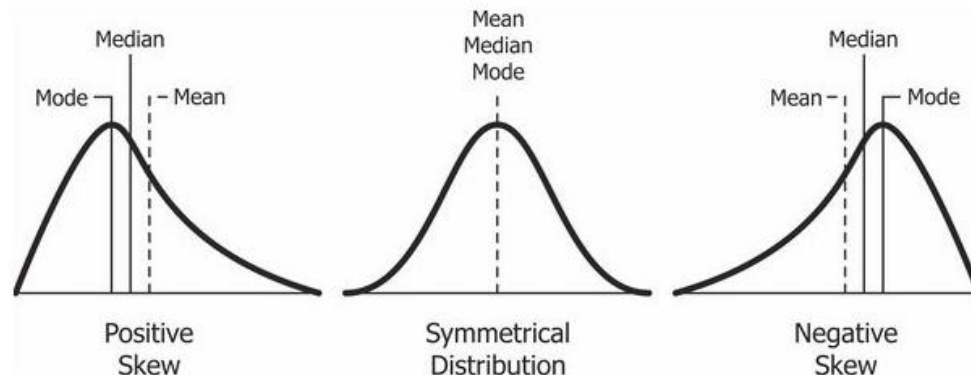


```
> tapply(y,Treatment,skewness)
```

Assumptions

A	B	C	D
1.744465	2.728535	2.320537	2.362395

- **Skewness:** It is the degree of distortion from the symmetrical bell curve or the normal distribution.
 - A symmetrical distribution will have a skewness of 0.
 - if the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
 - If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
 - If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.



Assumptions

- **Kurtosis**: It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution
 - **High kurtosis** in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why do we have so many outliers. **Kurtosis > 3**
 - **Low kurtosis** in a data set is an indicator that data has light tails or lack of outliers. **Kurtosis < 3**

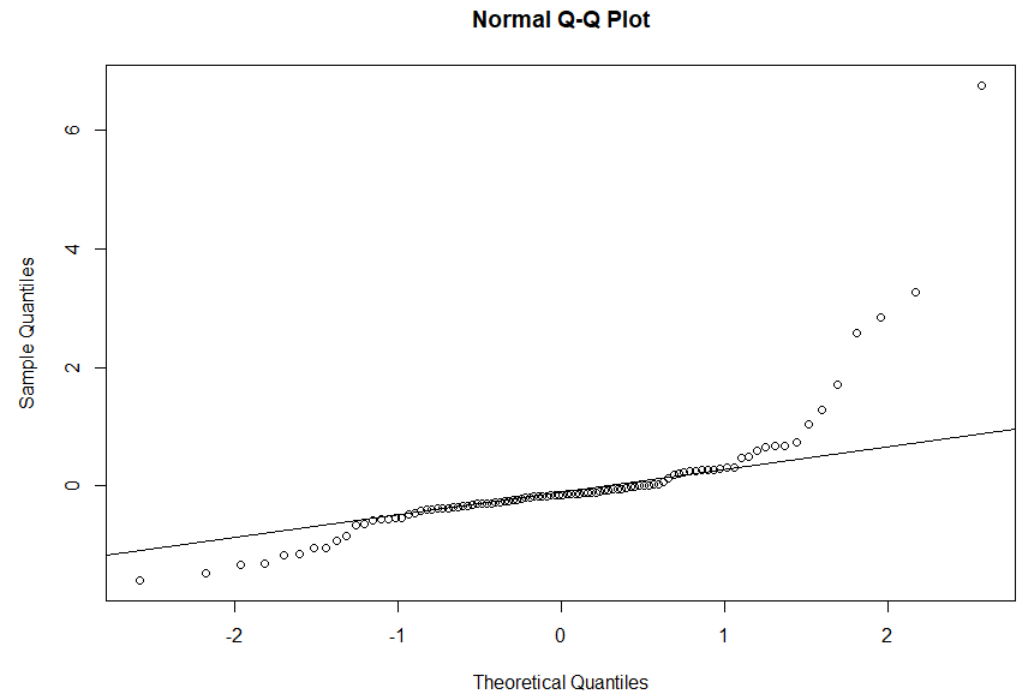
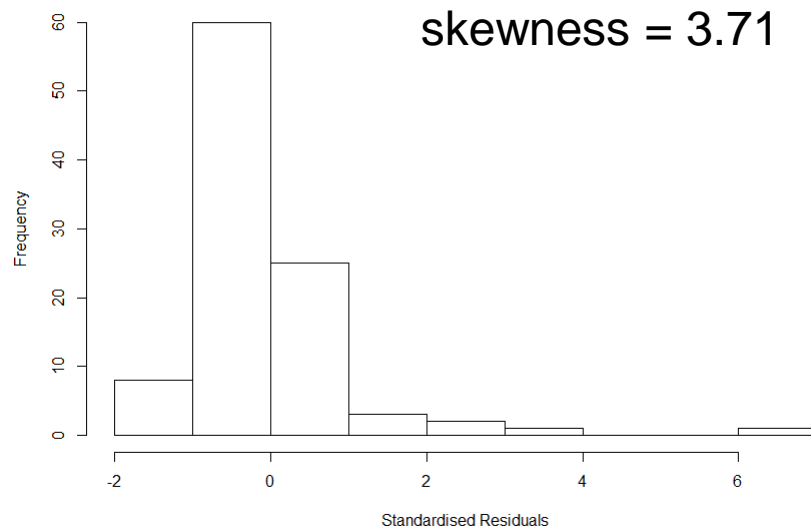
```
> tapply(y,Treatment,kurtosis)
```

A	B	C	D
4.807209	11.564391	8.460269	9.181076

Assumptions

Normality

```
> y.aov <- aov(y ~ Treatment)
> hist(rstandard(y.aov))
> qqnorm(rstandard(y.aov))
> qqline(rstandard(y.aov))
```



Assumptions: Normality

- **Shapiro-Wilk Test for normality** tests the null hypothesis that the data is Normally distributed:
 - H_0 : the data is consistent with Normal distribution
 - H_1 : the data is not consistent with Normal distribution
- Note: this test is very sensitive to sample size, particularly small sample size. It is recommended to use this test in conjunction with graphical methods.
 - Some people only use graphical methods.

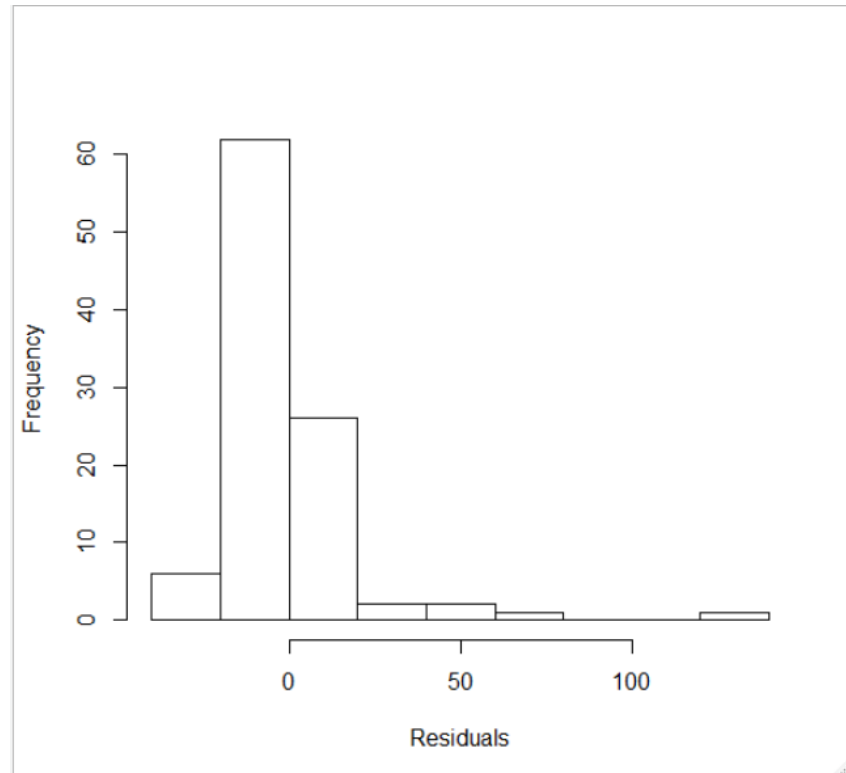
Assumptions: Normality

```
> shapiro.test(resid(y.aov))
```

Shapiro-wilk normality test

data: resid(y.aov)

W = 0.66003, p-value = 6.934e-14



Assumptions

Constant variance – ideal case

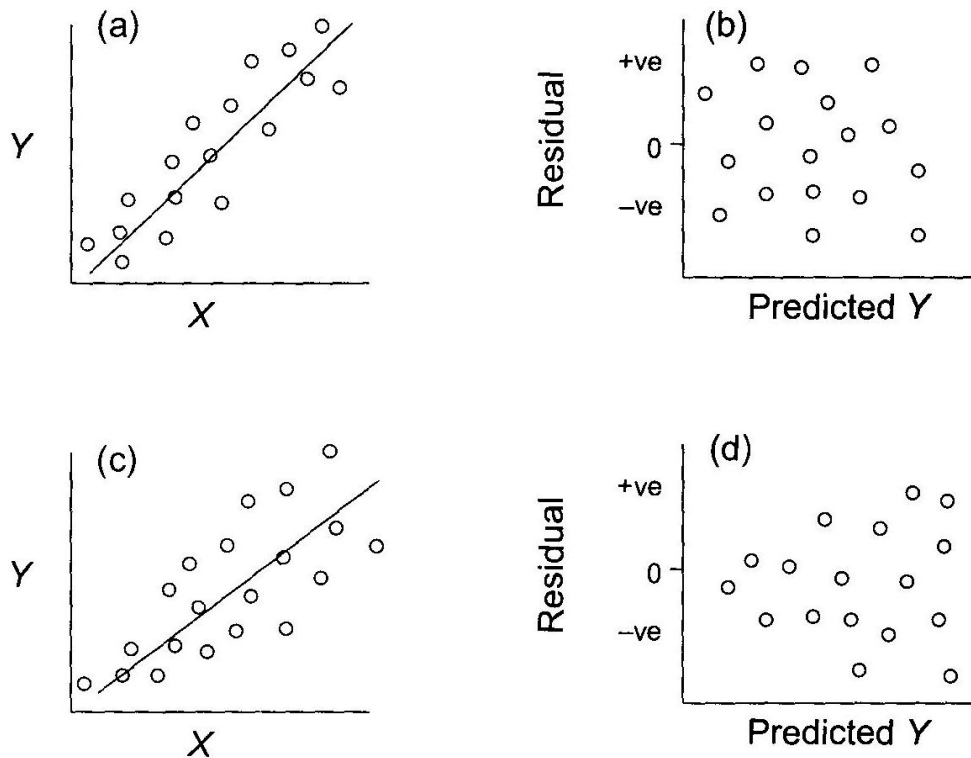


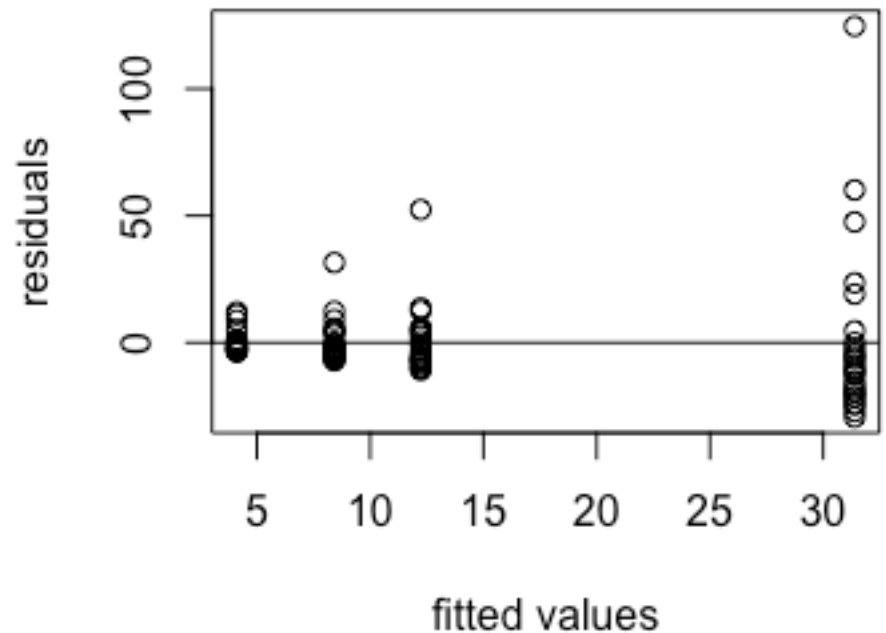
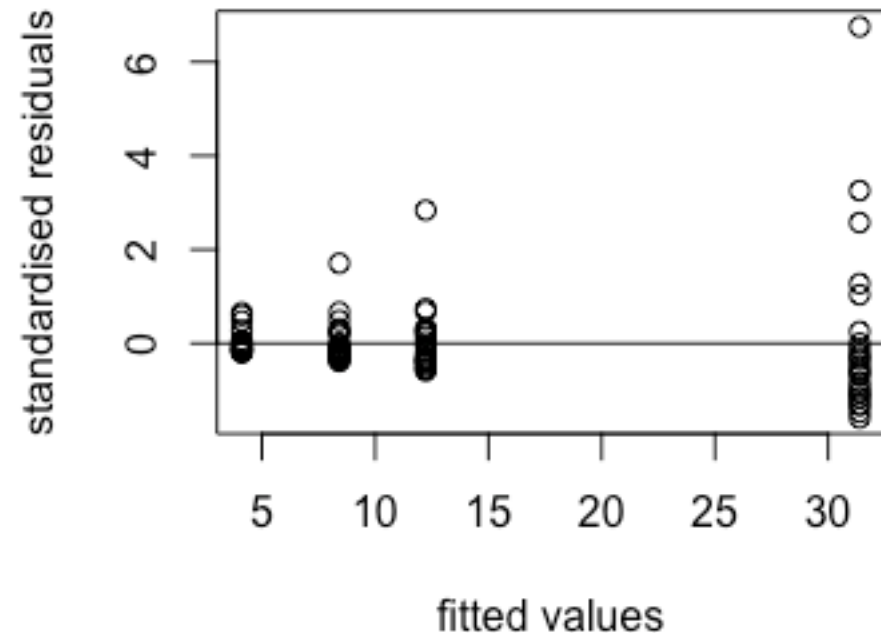
Figure 5.10 Diagrammatic representation of residual plots from linear regression: (a) regression showing even spread around line, (b) associated residual plot, (c) regression showing increasing spread around line, and (d) associated residual plot showing characteristic wedge-shape typical of skewed distribution.

p. 97 (Quinn & Keough, 2002)

Assumptions

Constant variance

```
> y.aov <- aov(y ~ Treatment)
> plot(fitted(y.aov), rstandard(y.aov))
> abline(0,0)
> plot(fitted(y.aov), resid(y.aov))
> abline(0,0)
```



```
> tapply(y, Treatment, sd)
      A      B      C      D
4.457697 13.027553 34.078634  8.354337
```

Unequal variance test

› Hypotheses

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

› Test statistic

$$\text{variance ratio} = \frac{\text{larger } s^2}{\text{smaller } s^2}$$

- › If test statistic > critical value, reject null hypothesis (H_0)
- › Critical value obtained from

Table of F_{crit} values

$$df = n_1 - 1, n_2 - 1$$

Equal variance test

```
> bartlett.test(y ~ Treatment)
```

Bartlett test of homogeneity of variances

data: y by Treatment

Bartlett's K-squared = 95.877, df = 3, p-value < 2.2e-16

Equal variance test

Levene's Test (F-Test) for equal spread (only for two variances)

e.g.

```
> var.test(heart.rate ~ Treatment, data = drink)
```

F test to compare two variances

data: heart.rate by Treatment

$F = 1.1842$, num df = 11, denom df = 11, p-value = 0.7841

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.3409134 4.1136607

sample estimates:

ratio of variances

1.184231

Data transformation

- › Unless the data is normally distributed we cannot calculate CI's, work out probabilities etc
- › We need to transform the data
- › Then repeat diagnostics on transformed data
- › And analyse the data on the transformed scale
- › Finally, we may **back-transform** the data to aid in interpretation

Data transformation

› Negatively skewed data with zero's

- $y' = \log_{10} y + C$
- C is a constant added to the data so that the smallest value is 1
 - Often $C = 1$

› Moderately positively skewed data

- $y' = \sqrt{k-y}$
- k is a constant added to the data so that the smallest score = 1

› Strongly positively skewed data

- $y' = \log(k-y)$
- k is a constant added to the data so that the smallest score = 1

› Proportions

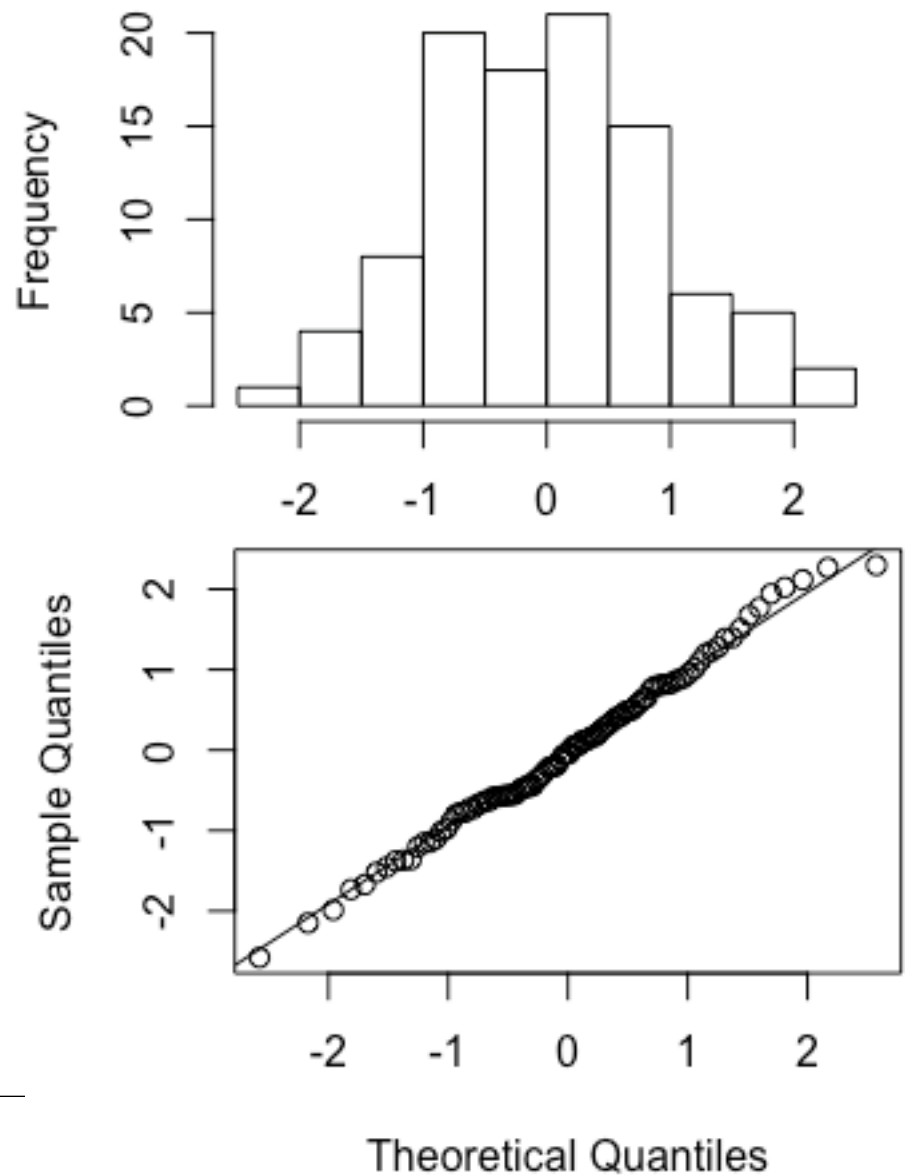
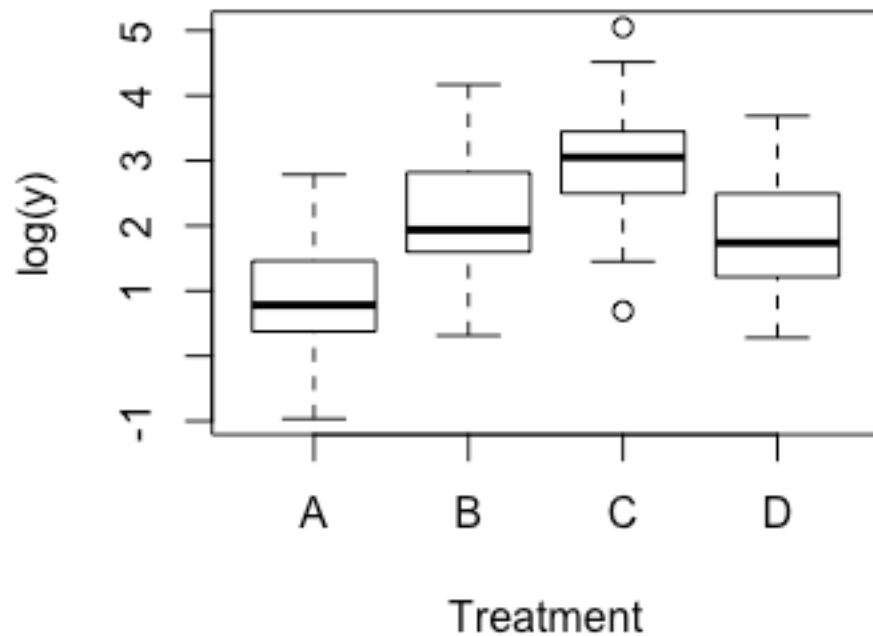
- $y' = \text{ArcSin}(y)$

Assumptions

```
y.aov <- aov(log(y) ~ Treatment)
```

Normality

- Log transform data



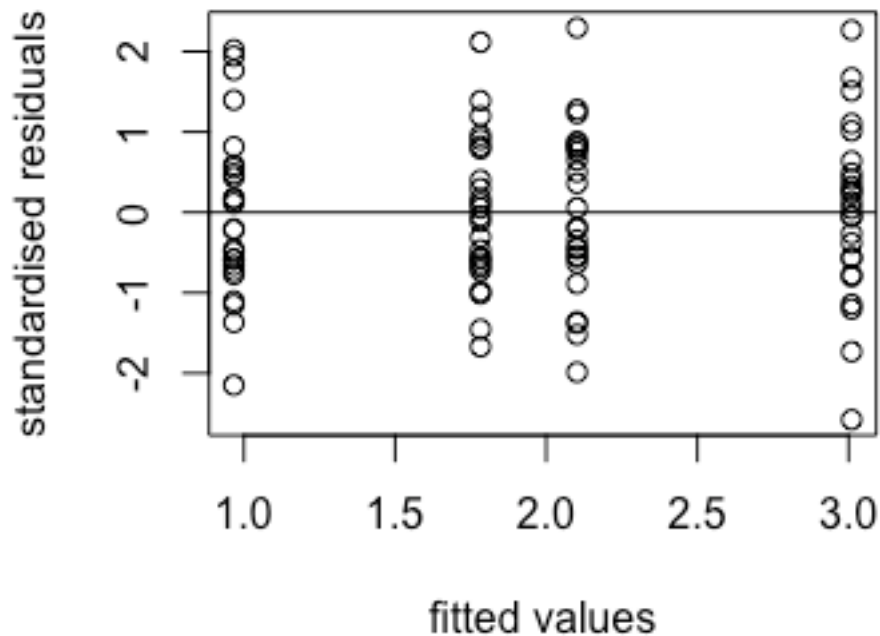
Assumptions

```
y.aov <- aov(log(y) ~ Treatment)
```

Constant variance

- Log transform data

```
> tapply(y, Treatment, sd)  
      A      B      C      D  
0.9467324 0.9188734 0.9724214 0.8304201
```



However

Ecology, 92(1), 2011, pp. 3–10
© 2011 by the Ecological Society of America

The arcsine is asinine: the analysis of proportions in ecology

DAVID I. WARTON^{1,2,3} AND FRANCIS K. C. HUI¹

¹*School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW 2052 Australia*

²*Evolution and Ecology Research Centre, The University of New South Wales, Sydney, NSW 2052 Australia*

Received: 22 May 2017 | Revised: 11 November 2017 | Accepted: 6 December 2017

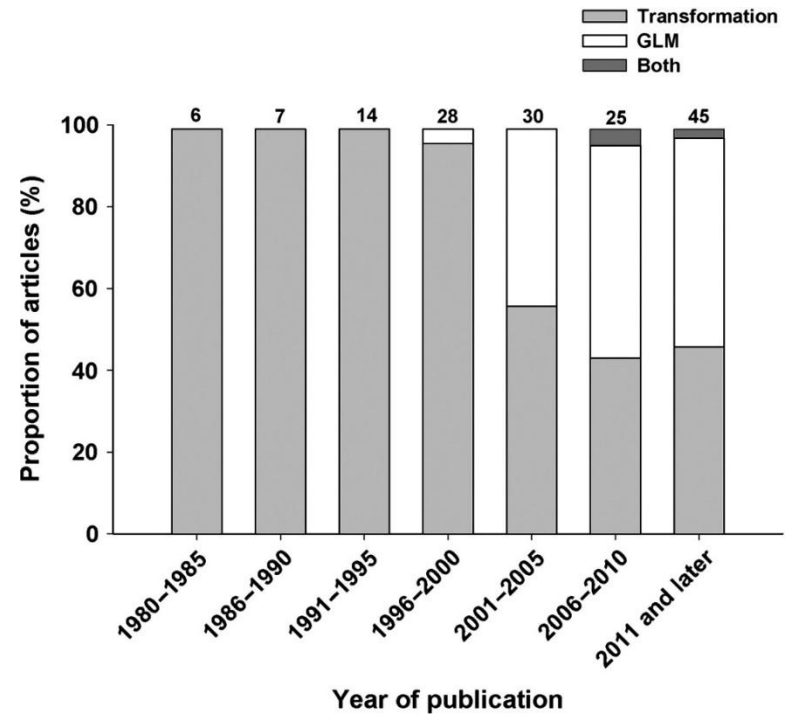
DOI: 10.1002/ece3.3807

ORIGINAL RESEARCH

WILEY *Ecology and Evolution*

Count data in biology—Data transformation or model reformation?

Anne P. St-Pierre¹ | Violaine Shikon² | David C. Schneider¹



Non-parametric alternative

Kruskal-Wallis test

- A non-parametric version of a one-way ANOVA test
 - Rank-based test that is an extension to Mann-Whitney test
- Compare medians of two or more groups
- Assumptions: the distributions of the groups is similar (except the medians).

```
> model.kruskal <- kruskal.test(WtGain ~ Diet, data = chicks)
```

```
➤ model.kruskal
```

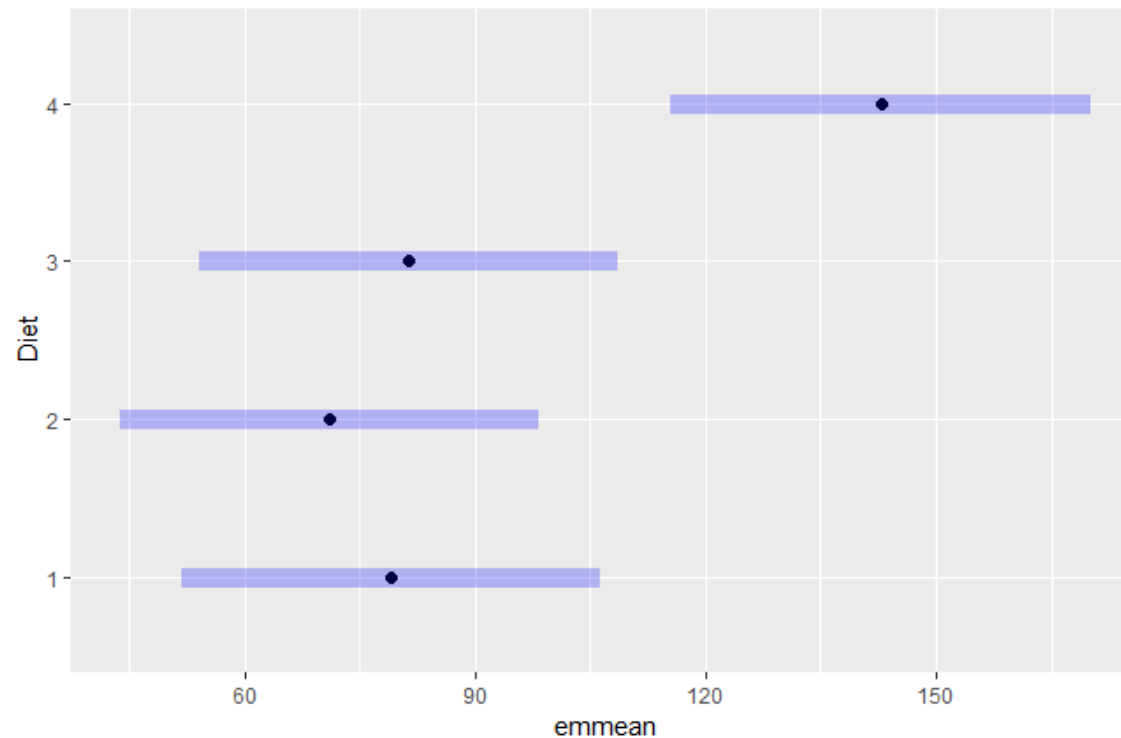
```
➤ Kruskal-wallis rank sum test
```

```
➤ data: WtGain by Diet
```

```
➤ Kruskal-wallis chi-squared = 7.0625, df = 3, p-value = 0.06993
```

Significance between pairs of treatments (groups)

Part 2



Significance between pairs of treatments (groups)

We have a significant F-test from the ANOVA: which means differ?

- We now wish to compare pairs of treatment means to see which are significantly different.
- To compare two groups we used a t-test
 - $t = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{s^2 \left(\frac{1}{n_2} + \frac{1}{n_1} \right)}} = \frac{\bar{y}_2 - \bar{y}_1}{se(\bar{y}_2 - \bar{y}_1)} = \frac{\text{difference in mean}}{\text{standard error of the difference in mean}}$
 - For ANOVA we use $df = \text{residual df}$ & $\text{Residual MS} = s^2$

Source	df	Sums-of-square (SS)	Mean-square (MS)	F statistic
Treatment	3	16,467	5,489	6.65
Residual	16	13,212	826	
Total	19	29,679		

Significant between pairs of treatments (groups)

Chicks example

– Diet 1 vs Diet 2

$$- t = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\text{ResMS}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{79 - 71}{\sqrt{826\left(\frac{1}{5} + \frac{1}{5}\right)}} = \frac{8}{18.18} = 0.44$$

– $df = 16, P = 0.67$

– no significant difference

Source	df	Sums-of-square (SS)	Mean-square (MS)	F statistic
Treatment	3	16,467	5,489	6.65
Residual	16	13,212	826	
Total	19	29,679		

Significant between pairs of treatments (groups)

Chicks example

- Using $\bar{y}_1 = 79.0$, $\bar{y}_2 = 71.0$, $\bar{y}_3 = 81.4$, $\bar{y}_4 = 142.8$, for other comparisons:

Diet 1 vs Diet 3:	$t = 0.42$	$P = 0.68$	\Rightarrow not significant
Diet 1 vs Diet 4:	$t = -3.51$	$P = 0.003$	\Rightarrow significant
Diet 2 vs Diet 3:	$t = -0.02$	$P = 0.98$	\Rightarrow not significant
Diet 2 vs Diet 4:	$t = -3.95$	$P = 0.001$	\Rightarrow significant
Diet 3 vs Diet 4:	$t = -3.93$	$P = 0.001$	\Rightarrow significant
- So Diet 4 mean has a significantly higher mean than Diets 1-3.

Significance between pairs of treatments (groups)

We have a significant F-test from the ANOVA: which means differ?

- We could look at the 95% CI for each mean and see which overlap

```
> emmeans(model.aov, "Diet")
```

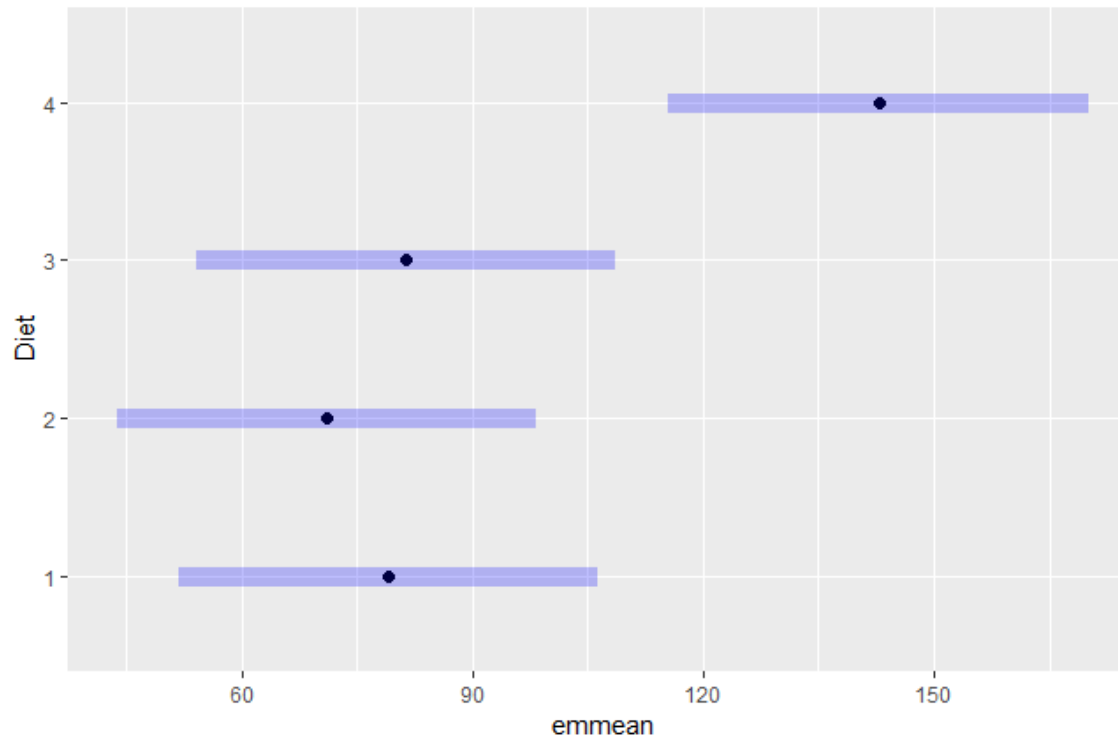
Diet	emmean	SE	df	lower.CL	upper.CL
1	79.0	12.85107	16	51.75695	106.24305
2	71.0	12.85107	16	43.75695	98.24305
3	81.4	12.85107	16	54.15695	108.64305
4	142.8	12.85107	16	115.55695	170.04305

Significance between pairs of treatments (groups)

We have a significant F-test from the ANOVA: which means differ?

- We could look at the 95% CI for each mean and see which overlap

```
> plot(emmeans(model.aov, "Diet"))
```

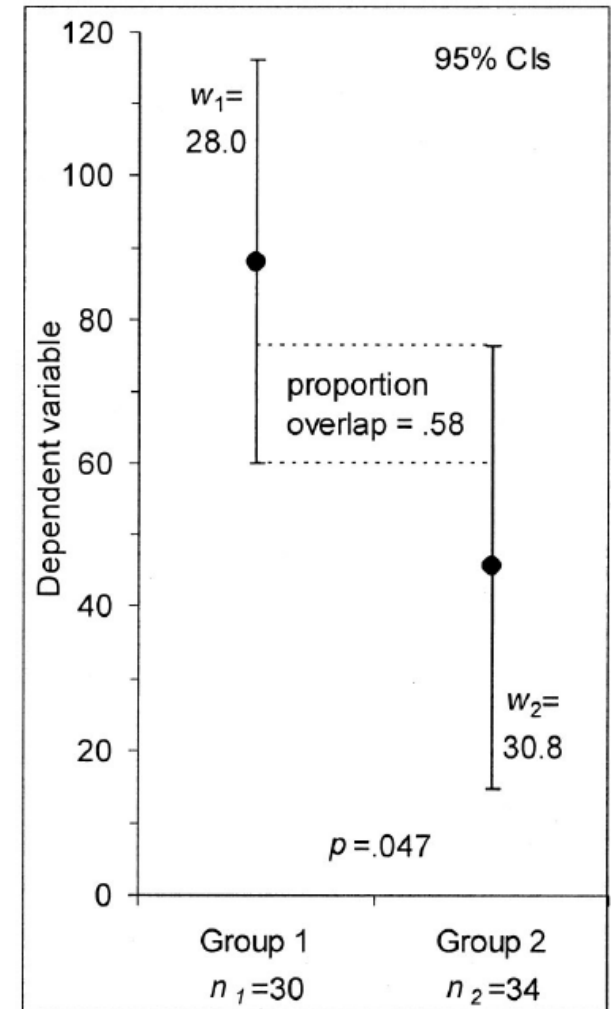


Significance between pairs of treatments (groups)

Overlapping 95% CIs

- Overlapping 95% CI:
 - $se(\bar{y}_1) + se(\bar{y}_2)$
- Standard error of difference between means
 - $\sqrt{se(\bar{y}_1)^2 + se(\bar{y}_2)^2}$
- SE difference < Overlapping CI always
- Touching 95% CI ~ P-value < 0.01
- 95% CI has 50% overlap relative to w_d then P-value < 0.05
 - W_d is margin error = 0.5 width 95% CI

Cumming, G. and S. Finch (2005). "Inference by eye: confidence intervals and how to read pictures of data." American Psychologist 60(2): 170.



Significant between pairs of treatments (groups)

Experiment-wise, family-wise or family error rates

- For every hypothesis test you perform, there is a type I error rate, which your significance level (α) defines.
 - In other words, there's a chance that you'll reject a null hypothesis that is actually true—it's a false positive.
 - Only one test, the type I error rate equals your significance level, which is often 5% (individual error rate)
- If you perform enough tests, you're virtually guaranteed to get a false positive!
- Family error rate is the overall error rate for all tests.

Significant between pairs of treatments (groups)

Experiment-wise, family-wise or family error rates

All Pairwise Comparisons Alpha = 0.05		
Groups	Comparisons	Experimentwise Error Rate
2	1	0.05
3	3	0.142625
4	6	0.264908109
5	10	0.401263061
6	15	0.53670877
7	21	0.659438374
8	28	0.762173115
9	36	0.842220785
10	45	0.900559743
11	55	0.940461445
12	66	0.966134464
13	78	0.981700416
14	91	0.990606054
15	105	0.995418807
From StatisticsByJim.com		

$$\text{family error} = 1 - (1 - \alpha)^c$$

Where:

α = significance level

c = number of comparisons

<https://statisticsbyjim.com/anova/post-hoc-tests-anova/>

Experiment-wise, family-wise or family error rates

How to control type 1 family errors:

- Only look at pre-defined comparisons *a priori*.
- Use graphical methods only
- Adjust post-hoc p-value to reduce type 1 error
- E.g. Bonferroni Correction: $\frac{\alpha}{c}$

Where: α = significance level

c = number of comparisons

- Or set $\alpha = 0.01$
- Set the family error rate for post-hoc tests
 - E.g. Tukey test in R

Experiment-wise, family-wise or family error rates

Post Hoc Tests and the Statistical Power Trade-off

- Post hoc tests control the experiment-wise error rate by reducing the statistical power of the comparisons.
- To obtain a lower family error rate, the procedures must lower the significance level for all individual comparisons.
- As the number of comparisons increases, the post hoc analysis is forced to lower the individual significance level even further.
 - E.g. with 4 comparisons above, you may end up setting each individual pair-wise test at $\alpha = 0.013$ to keep family error rate at 0.05.

Significant between pairs of treatments (groups)

Post-hoc tests in R

- Tukey's method using the base R function `TukeyHSD`
- `emmeans(aov.model, pairwise ~ group)` from `emmeans` package

```
> TukeyHSD(y.aov)
```

```
Tukey multiple comparisons of means  
95% family-wise confidence level
```

```
Fit: aov(formula = y ~ Treatment)
```

```
$Treatment
```

	diff	lwr	upr	p adj
B-A	8.120435	-5.816849	22.057718	0.4276485
C-A	27.273217	13.335934	41.210501	0.0000094
D-A	4.304475	-9.632808	18.241759	0.8507622
C-B	19.152782	5.215499	33.090066	0.0028684
D-B	-3.815960	-17.753243	10.121324	0.8905719
D-C	-22.968742	-36.906026	-9.031459	0.0002294

Part 3: Example

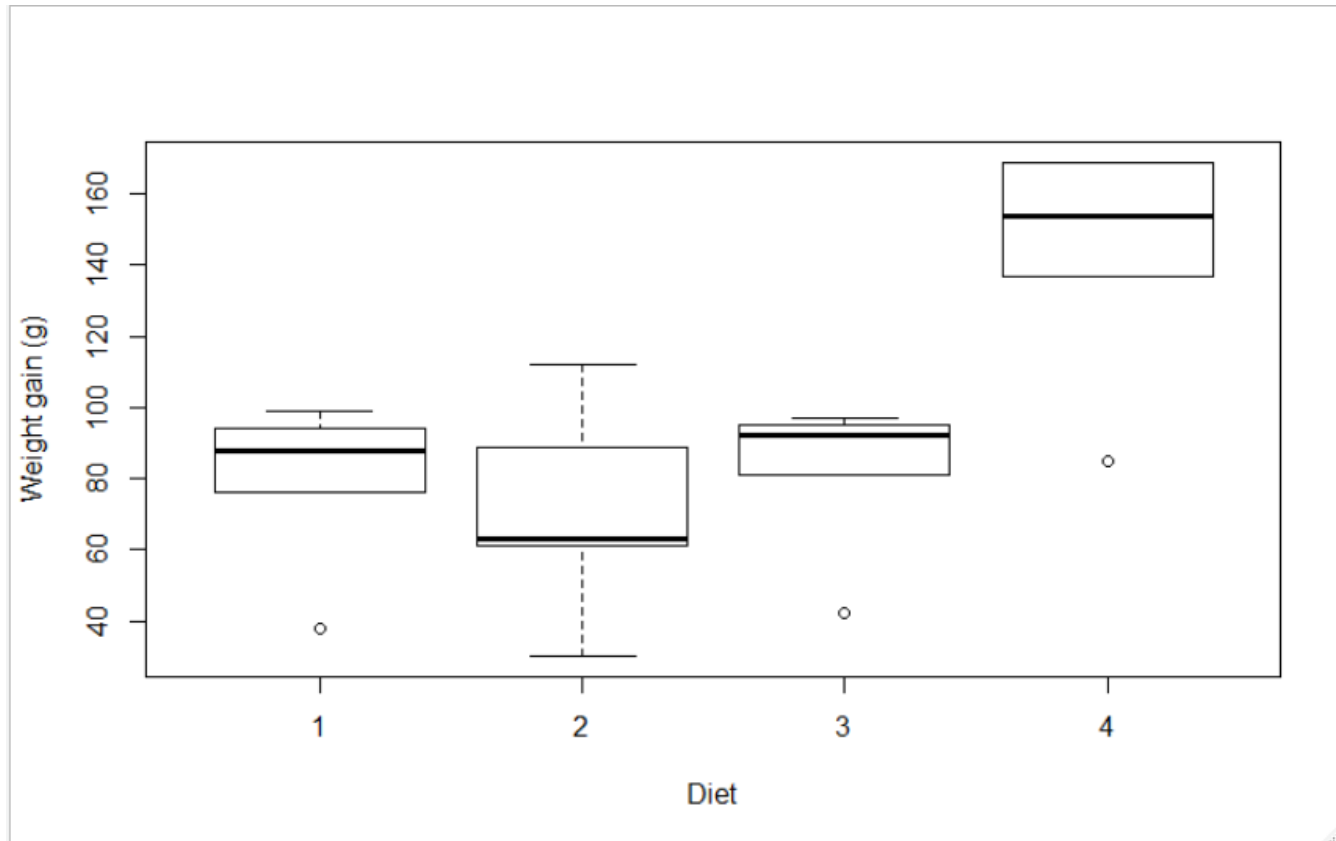
- A study was undertaken to compare the weight gains of chicks on four different diets.
- Twenty similar chicks were used in the study and were randomly allocated to one of the four groups.
- The allocation was done in such a way as to have equal replication (five chicks) in each treatment group.



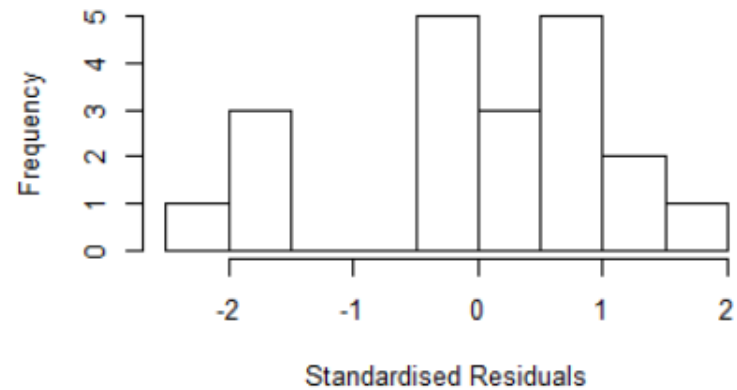
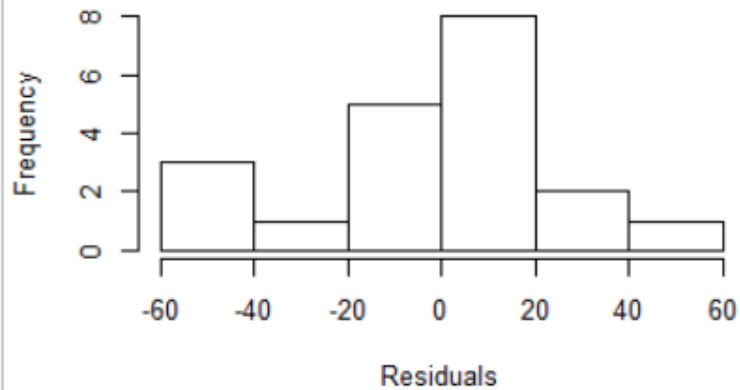
Significant between pairs of treatments (groups)



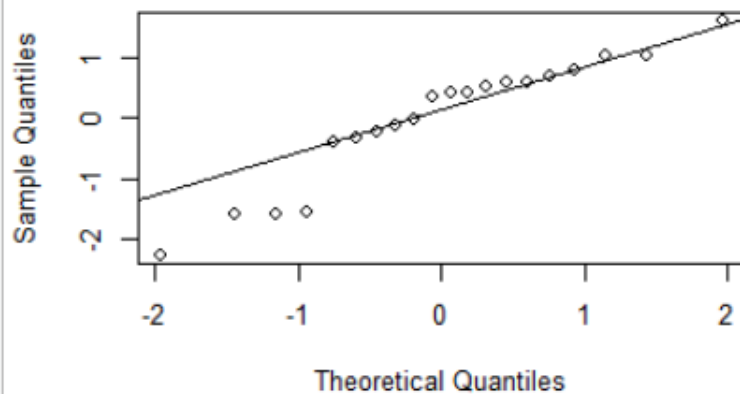
Chicks example



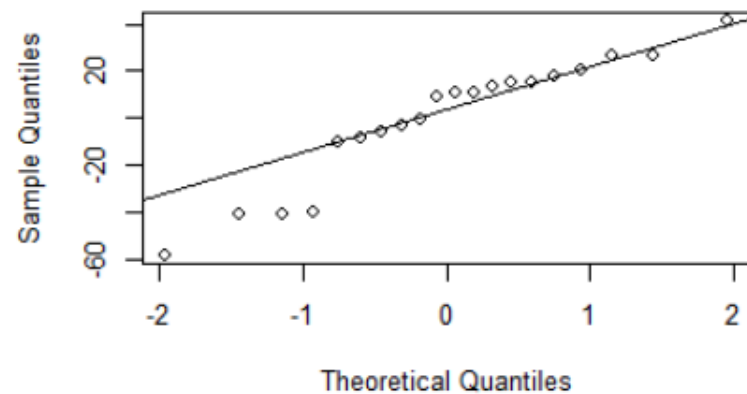
Assumptions



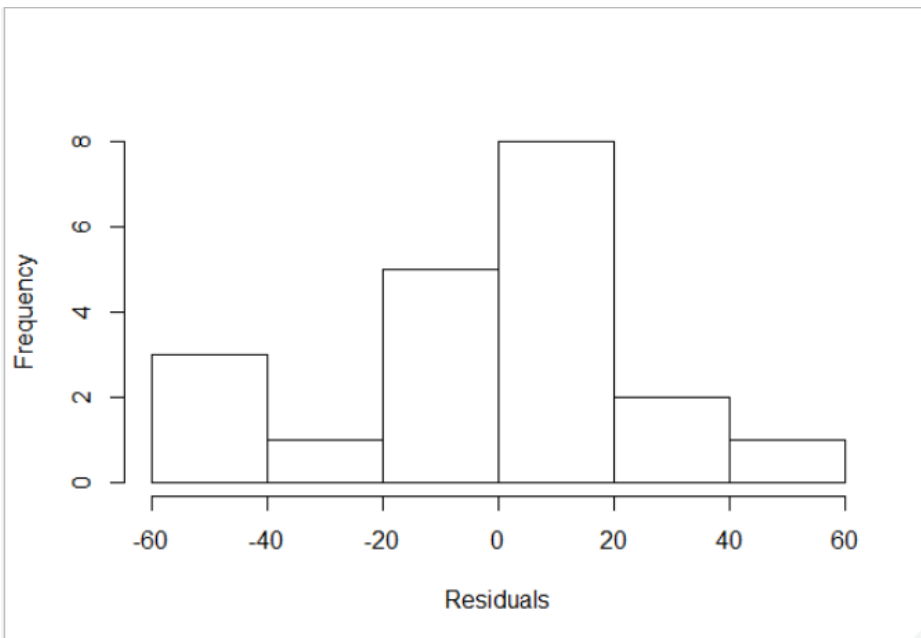
Normal Q-Q Plot: Standardised Residuals



Normal Q-Q Plot: Residuals



Assumptions



```
> shapiro.test(resid(model.aov))
```

Shapiro-wilk normality test

```
data: resid(model.aov)
```

```
w = 0.90961, p-value = 0.06265
```

Assumptions

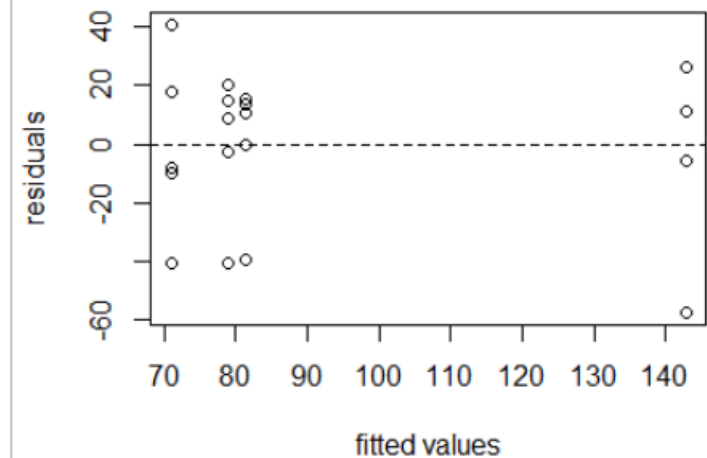
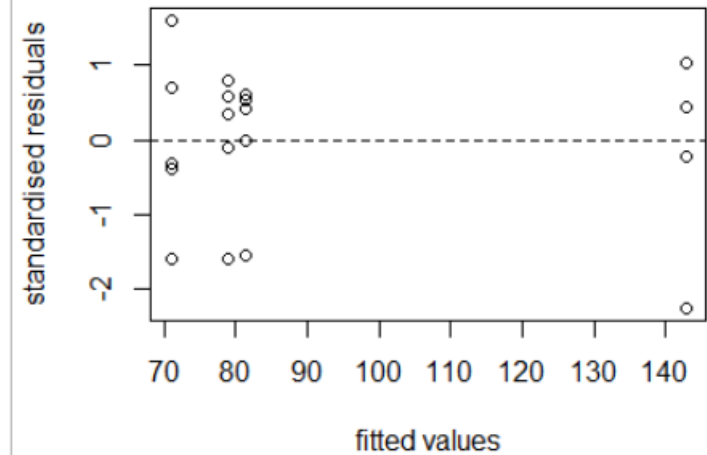
Equal variance

```
> bartlett.test(WtGain ~ Diet, data = chicks)
```

Bartlett test of homogeneity of variances

data: WtGain by Diet

Bartlett's χ^2 = 0.85164, df = 3, p-value = 0.8371



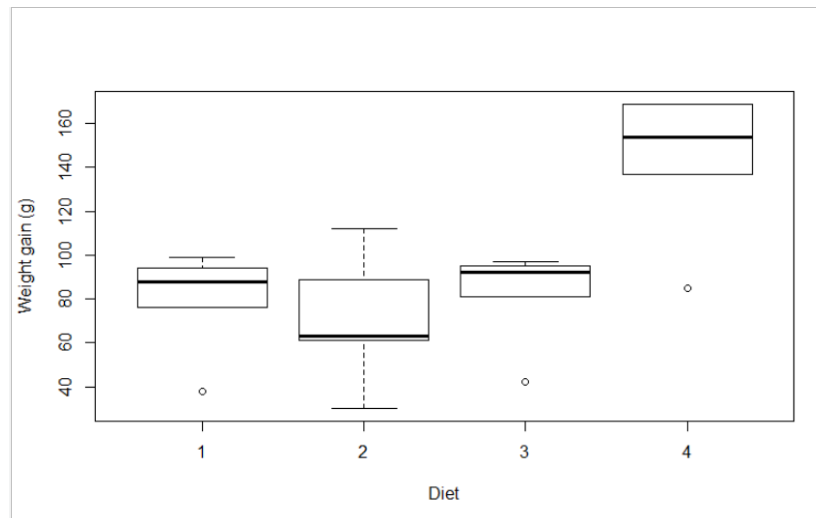
Results



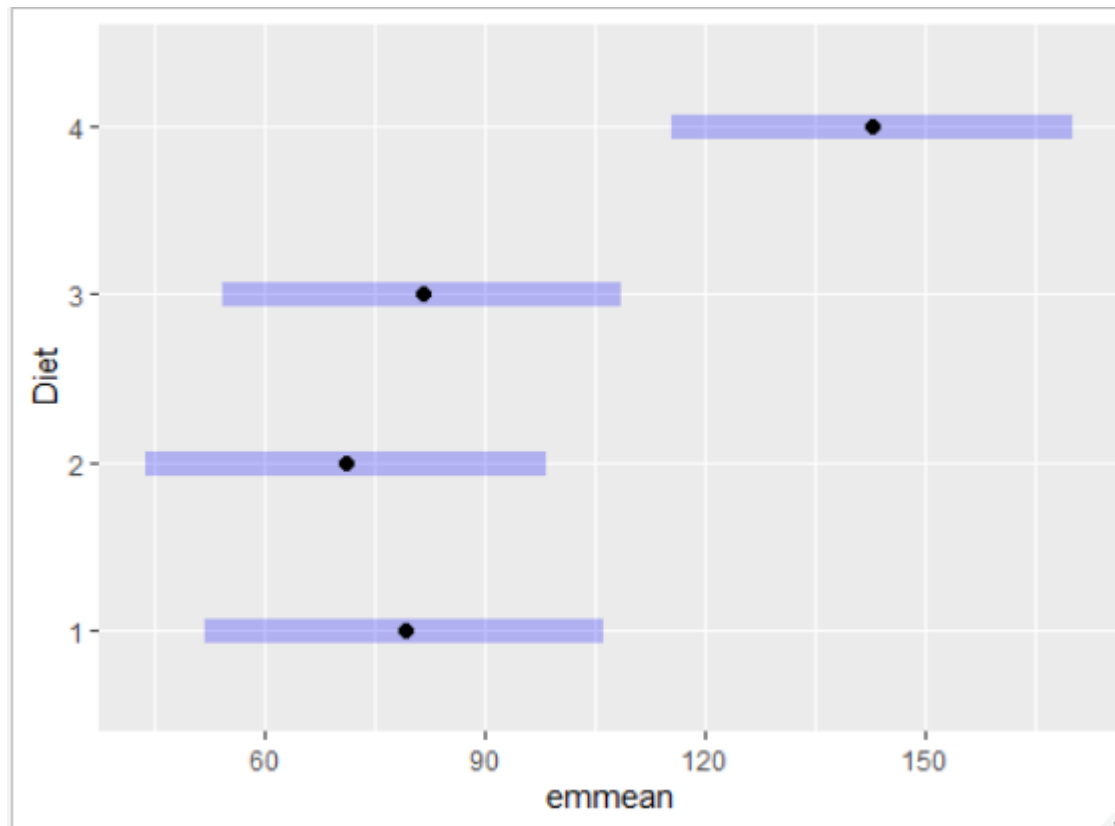
```
> model.aov <- aov(WtGain ~ Diet, data = chicks)
> summary(model.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	3	16467	5489	6.647	0.004 **
Residuals	16	13212	826		

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Post-hoc tests



Post-hoc tests

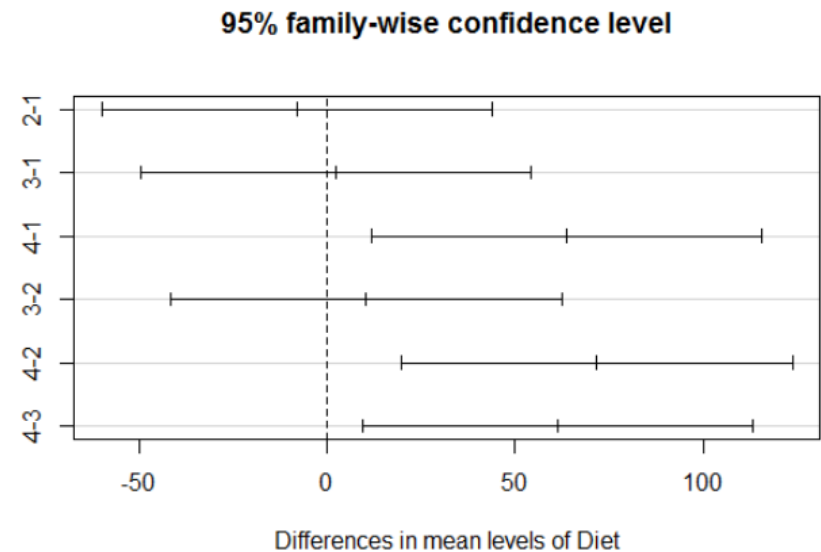


```
> TukeyHSD(model.aov)
```

Tukey multiple comparisons of means
95% family-wise confidence level

```
Fit: aov(formula = wtGain ~ Diet, data = chicks)
```

\$Diet	diff	lwr	upr	p adj
2-1	-8.0	-59.996625	43.99662	0.9705671
3-1	2.4	-49.596625	54.39662	0.9991415
4-1	63.8	11.803375	115.79662	0.0138464
3-2	10.4	-41.596625	62.39662	0.9389332
4-2	71.8	19.803375	123.79662	0.0056645
4-3	61.4	9.403375	113.39662	0.0180630





Post-hoc tests

```
> emmeans(model.aov, pairwise ~ Diet)
```

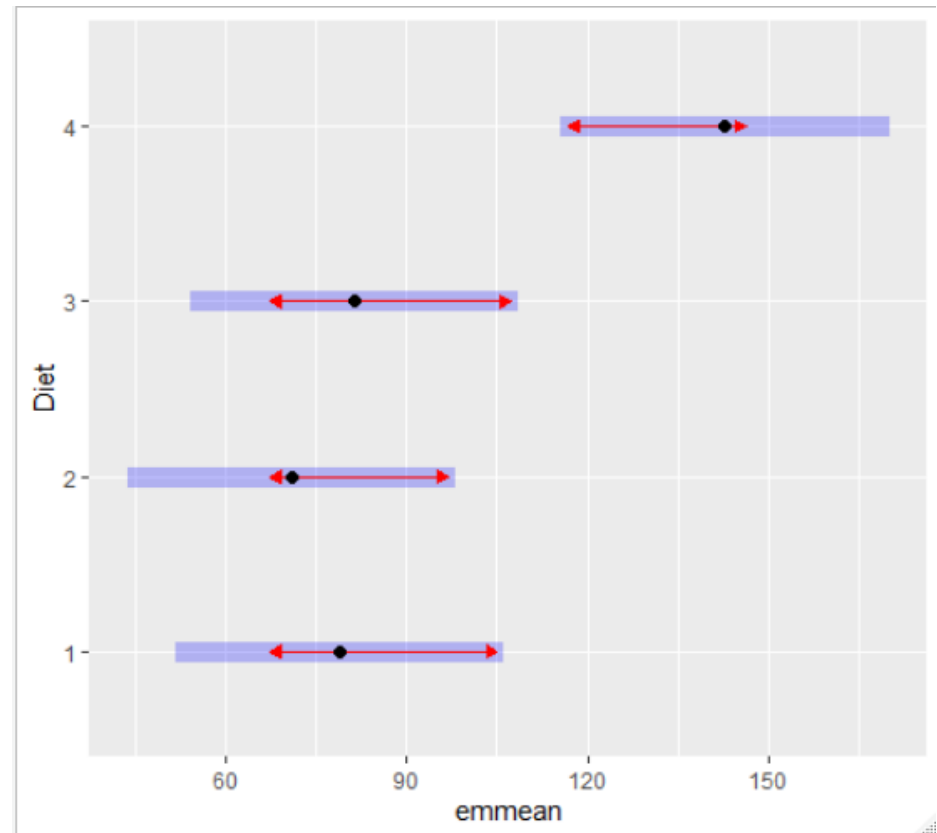
```
$emmeans
```

Diet	emmean	SE	df	lower.CL	upper.CL
1	79.0	12.9	16	51.8	106.2
2	71.0	12.9	16	43.8	98.2
3	81.4	12.9	16	54.2	108.6
4	142.8	12.9	16	115.6	170.0

Confidence level used: 0.95

```
$contrasts
```

contrast	estimate	SE	df	t.ratio	p.value
1 - 2	8.0	18.2	16	0.440	0.9706
1 - 3	-2.4	18.2	16	-0.132	0.9991
1 - 4	-63.8	18.2	16	-3.510	0.0138
2 - 3	-10.4	18.2	16	-0.572	0.9389
2 - 4	-71.8	18.2	16	-3.951	0.0057
3 - 4	-61.4	18.2	16	-3.378	0.0181



P value adjustment: tukey method for comparing a family of 4 estimates

Conclusions

Assumptions

- Use residuals to test assumptions for your models e.g. ANOVA
 - Graphical and formal tests available

Pairwise Tests

- Graphical methods
- Formal tests: Tukey's test
- Some propose to only look at pre-determined comparisons
- For multiple tests, the chance of at least one false result is increased.
 - this is the family error rate
- Depends on your statistical church...

Resources

- Quinn & Keough (2002) or (2024). [Experimental Design and Data Analysis for Biologists](#)
- Fox, G. A., S. Negrete-Yankelevich, and V. J. Sosa. (2015). [Ecological statistics: contemporary theory and application](#). Oxford University Press, USA.
- Logan (2010) [Biostatistical design and analysis using R a practical guide](#)
- Ebooks in library

– **Questions/Feedback?**



THE UNIVERSITY OF
SYDNEY