

Basic Statistics

A primer in basic statistics for BCB (Hons) 2019

AJ Smit and Robert Schlegel

2019-04-09

Contents

1	Introduction	9
1.1	Venue, date and time	9
1.2	Course outline	9
1.3	About this Workshop	9
1.4	This is biology: why more R coding?	9
1.5	Installing R and RStudio	9
1.6	Resources	9
1.7	Style and code conventions	9
1.8	Assessment and teaching philosophy	9
1.9	About this document	9
2	Types of data	10
2.1	Data classes	11
2.1.1	Numerical data	11
2.1.2	Qualitative data	11
2.1.3	Binary data	11
2.1.4	Character values	11
2.1.5	Missing values	11
2.1.6	Complex numbers	11
2.2	Viewing our data	11
2.2.1	From the Environment pane	11
2.2.2	head() and tail()	11
2.2.3	colnames()	11
2.2.4	summary()	11
3	Descriptive statistics: central tendency and dispersion	12
3.1	Samples and populations	13
3.2	Measures of central tendency	13
3.2.1	The mean	13
3.2.2	The median	13
3.2.3	Skewness	13
3.2.4	Kurtosis	13
3.3	Measures of variation and spread	13
3.3.1	The variance and standard deviation	13
3.3.2	Quantiles	13
3.3.3	The minimum, maximum and range	13
3.3.4	Covariance	13
3.3.5	Correlation	13
3.4	Missing values	13
3.5	Descriptive statistics by group	13
3.5.1	Groupwise summary statistics	13

3.5.2	Displays of group summaries	13
3.6	Exercises	13
3.6.1	Exercise 1	13
4	Graphical data displays	14
4.1	Qualitative data	14
4.2	Continuous data	14
4.2.1	Frequency distributions (histograms)	14
4.2.2	Box plots	14
4.2.3	Pairwise Scatter plots	14
4.2.4	Bar graphs	14
4.2.5	Density graphs	14
4.2.6	Violin plots	14
4.3	Exercises	14
4.3.1	Exercise 1	14
5	Distributions	15
5.1	Discrete distributions	16
5.1.1	Bernoulli distribution	16
5.1.2	Binomial distribution	16
5.1.3	Negative binomial distribution	16
5.1.4	Geometric distribution	16
5.1.5	Poisson distribution	16
5.2	Continuous distributions	16
5.2.1	Normal distribution	16
5.2.2	Uniform distribution	16
5.2.3	Student T distribution	16
5.2.4	Chi-squared distribution	16
5.2.5	Exponential distribution	16
5.2.6	F distribution	16
5.2.7	Gamma distribution	16
5.2.8	Beta distribution	16
5.2.9	Paranormal distributions	16
5.3	Finding one's data distribution	16
5.4	Exercises	16
5.4.1	Exercise 1	16
6	Inferences about one or two populations	17
6.1	Assumptions	19
6.1.1	Normality	19
6.1.2	Homoscedasticity	19
6.1.3	Two for one	19
6.2	One-sample t -tests	19
6.2.1	One-sided one-sample t -tests	19
6.2.2	Two-sided one-sample t -tests	19
6.3	Two-sample t -tests	19
6.3.1	One-sided two-sample t -tests	19
6.3.2	Two-sided two-sample t -tests	19
6.4	Paired t -tests	19
6.5	Comparison of two population proportions	19
6.5.1	One-sample and two-sample tests	19
6.5.2	One-sided and two-sided tests	19

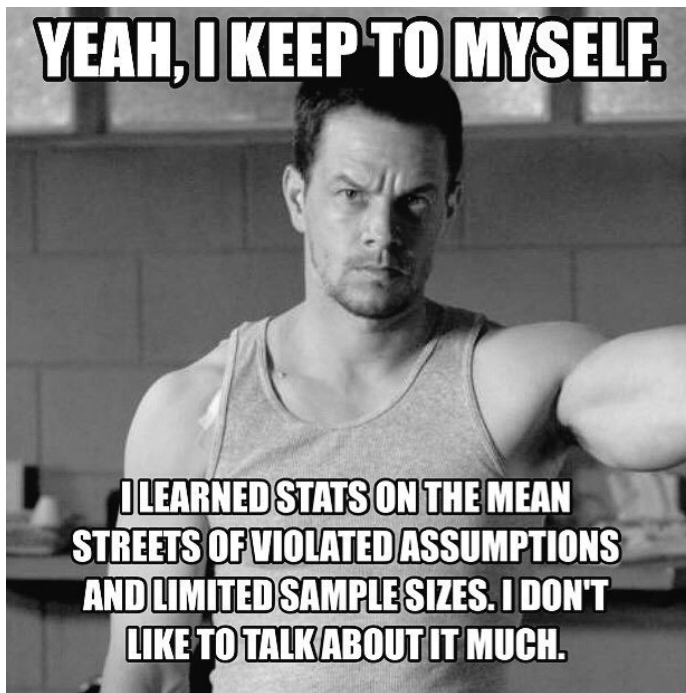
6.6	A <i>t</i> -test workflow	19
6.6.1	Loading data	19
6.6.2	Visualising data	19
6.6.3	Formulating a hypothesis	19
6.6.4	Choosing a test	19
6.6.5	Checking assumptions	19
6.6.6	Running an analysis	19
6.6.7	Interpreting the results	19
6.6.8	Drawing conclusions	19
6.6.9	Going further	19
6.7	Exercises	19
6.7.1	Exercise 1	19
6.7.2	Exercise 2	19
7	ANOVA	20
7.1	Remember the <i>t</i> -test	21
7.2	ANOVA	21
7.2.1	Single factor	21
7.2.2	Multiple factors	21
7.2.3	Examples	21
7.3	Alternatives to ANOVA	21
7.3.1	Wilcoxon rank sum test	21
7.3.2	Kruskal-Wallis rank sum test	21
7.3.3	The SA time data	21
7.4	Exercises	21
7.4.1	Exercise 1	21
7.4.2	Exercise 2	21
7.4.3	Exercise 3	21
8	Simple linear regressions	22
8.1	The simple linear regression equation	22
8.1.1	The intercept	22
8.1.2	The regression coefficient	22
8.1.3	A graph of the linear regression	22
8.1.4	Predicting from the linear model	22
8.1.5	The coefficient of determination, r^2	22
8.1.6	Significance test for linear regression	22
8.1.7	Confidence interval for linear regression	22
8.1.8	Prediction interval for linear regression	22
8.1.9	Residual plot	22
8.1.10	Standardised residual	22
8.1.11	Normal probability plot of residuals	22
8.2	Using an additional categorical variable	22
9	Correlations	23
9.1	Pearson correlation	23
9.2	Spearman rank correlation	23
9.3	Kendall rank correlation	23
9.4	One panel visual	23
9.5	Multiple panel visual	23
9.6	Exercises	23
9.6.1	Exercise 1	23

10 Confidence intervals	24
10.1 Calculating confidence	24
10.2 CI of compared means	24
10.3 Harrell plots	24
10.4 Exercises	24
10.4.1 Exercise 1	24
10.4.2 Exercise 2	24
11 Testing assumptions or: How I learned to stop worrying and transform the data	25
11.1 Backgroud	26
11.1.1 Normality	26
11.1.2 Homoscedasticity	26
11.1.3 Epic fail. Now what?	26
11.2 Transforming data	26
11.2.1 Log transform	26
11.2.2 Arcsine transform	26
11.2.3 Cube root	26
11.2.4 Square root transform	26
11.3 Exercises	26
11.3.1 Exercise 1	26
11.3.2 Exercise 2	26
12 Linear mixed models	27
12.1 Wilcox rank sum test	27
12.2 Kruskal-Wallis rank sum test	27
12.2.1 Single factor	27
12.2.2 Multiple factors	27
12.3 Generalised linear models	27
12.3.1 Sign Test	27
12.3.2 Wilcoxon Signed-Rank Test	27
12.3.3 Mann-Whitney-Wilcoxon Test	27
12.3.4 Kruskal-Wallis Test	27
12.3.5 Generalised linear models (GLM)	27
12.4 Exercises	27
12.5 Exercise 1	27
13 Chi-squared	28

List of Tables

List of Figures

Preface



This is a workshop about the practice of the basic statistics used by biologists, and not about the theory and mathematical underpinnings of the methods used. Each of the Chapters will cover a basic kind of statistical approach, and the main classes of data it applies to. Since much insight and understanding can be gained from visualising our data, we will also explore the main types of graphical summaries that best accompany the statistical methodologies. It is our intention to demonstrate how we go about analysing our data.

Prerequisites

A prerequisite for this course is a basic proficiency in using R (?). The necessary experience will have been gained from completing the Intro R Workshop: Data Manipulation, Analysis, and Graphing¹ Workshop that was part of your BCB Core Honours module (i.e. Biostatistics). You will also need a laptop with R and RStudio installed as per the instructions provided in that workshop. If you do not have a personal laptop, most computers in the 5th floor lab will be correctly set up for this purpose.

¹https://robwschlegel.github.io/Intro_R_Workshop/

1

Introduction

Placeholder

- 1.1 Venue, date and time**
- 1.2 Course outline**
- 1.3 About this Workshop**
- 1.4 This is biology: why more R coding?**
- 1.5 Installing R and RStudio**
- 1.6 Resources**
- 1.7 Style and code conventions**
- 1.8 Assessment and teaching philosophy**
- 1.9 About this document**

2

Types of data

Placeholder

2.1 Data classes

2.1.1 Numerical data

2.1.1.1 *Nominal (discrete) data*

2.1.1.2 *Continuous data*

2.1.1.3 *Dates*

2.1.2 Qualitative data

2.1.2.1 *Categorical data*

2.1.2.2 *Ordinal data*

2.1.3 Binary data

2.1.4 Character values

2.1.5 Missing values

2.1.6 Complex numbers

2.2 Viewing our data

2.2.1 From the Environment pane

2.2.2 `head()` and `tail()`

2.2.3 `colnames()`

2.2.4 `summary()`

3

Descriptive statistics: central tendency and dispersion

Placeholder

3.1 Samples and populations

3.2 Measures of central tendency

3.2.1 The mean

3.2.2 The median

3.2.3 Skewness

3.2.4 Kurtosis

3.3 Measures of variation and spread

3.3.1 The variance and standard deviation

3.3.2 Quantiles

3.3.3 The minimum, maximum and range

3.3.4 Covariance

3.3.5 Correlation

3.4 Missing values

3.5 Descriptive statistics by group

3.5.1 Groupwise summary statistics

3.5.2 Displays of group summaries

3.6 Exercises

3.6.1 Exercise 1

4

Graphical data displays

Placeholder

4.1 Qualitative data

4.2 Continuous data

4.2.1 Frequency distributions (histograms)

4.2.2 Box plots

4.2.3 Pairwise Scatter plots

4.2.4 Bar graphs

4.2.5 Density graphs

4.2.6 Violin plots

4.3 Exercises

4.3.1 Exercise 1

5

Distributions

Placeholder

5.1 Discrete distributions

5.1.1 Bernoulli distribution

5.1.2 Binomial distribution

5.1.3 Negative binomial distribution

5.1.4 Geometric distribution

5.1.5 Poisson distribution

5.2 Continuous distributions

5.2.1 Normal distribution

5.2.2 Uniform distribution

5.2.3 Student T distribution

5.2.4 Chi-squared distribution

5.2.5 Exponential distribution

5.2.6 F distribution

5.2.7 Gamma distribution

5.2.8 Beta distribution

5.2.9 Paranormal distributions

5.3 Finding one's data distribution

5.4 Exercises

5.4.1 Exercise 1

6

Inferences about one or two populations

Placeholder

6.1 Assumptions

6.1.1 Normality

6.1.2 Homoscedasticity

6.1.3 Two for one

6.2 One-sample t -tests

6.2.1 One-sided one-sample t -tests

6.2.2 Two-sided one-sample t -tests

6.3 Two-sample t -tests

6.3.1 One-sided two-sample t -tests

6.3.2 Two-sided two-sample t -tests

6.4 Paired t -tests

6.5 Comparison of two population proportions

6.5.1 One-sample and two-sample tests

6.5.2 One-sided and two-sided tests

6.6 A t -test workflow

6.6.1 Loading data

6.6.2 Visualising data

6.6.3 Formulating a hypothesis

6.6.4 Choosing a test

6.6.5 Checking assumptions

6.6.6 Running an analysis

6.6.7 Interpreting the results

6.6.8 Drawing conclusions

6.6.9 Going further

6.7 Exercises

6.7.1 Exercise 1

6.7.2 Exercise 2

7

ANOVA

Placeholder

7.1 Remember the t -test

7.2 ANOVA

7.2.1 Single factor

7.2.2 Multiple factors

7.2.2.1 *About interaction terms*

7.2.3 Examples

7.2.3.1 *Snakes!*

7.3 Alternatives to ANOVA

7.3.1 Wilcoxon rank sum test

7.3.2 Kruskal-Wallis rank sum test

7.3.2.1 *Single factor*

7.3.2.2 *Multiple factors*

7.3.3 The SA time data

7.4 Exercises

7.4.1 Exercise 1

7.4.2 Exercise 2

7.4.3 Exercise 3

8

Simple linear regressions

Placeholder

8.1 The simple linear regression equation

8.1.1 The intercept

8.1.2 The regression coefficient

8.1.3 A graph of the linear regression

8.1.4 Predicting from the linear model

8.1.5 The coefficient of determination, r^2

8.1.6 Significance test for linear regression

8.1.7 Confidence interval for linear regression

8.1.8 Prediction interval for linear regression

8.1.9 Residual plot

8.1.10 Standardised residual

8.1.11 Normal probability plot of residuals

8.2 Using an additional categorical variable

9

Correlations

Placeholder

- 9.1 Pearson correlation**
- 9.2 Spearman rank correlation**
- 9.3 Kendall rank correlation**
- 9.4 One panel visual**
- 9.5 Multiple panel visual**
- 9.6 Exercises**
 - 9.6.1 Exercise 1**

10

Confidence intervals

Placeholder

10.1 Calculating confidence

10.2 CI of compared means

10.3 Harrell plots

10.4 Exercises

10.4.1 Exercise 1

10.4.2 Exercise 2

11

Testing assumptions or: How I learned to stop worrying and transform the data

Placeholder

11.1 Background

11.1.1 Normality

11.1.2 Homoscedasticity

11.1.3 Epic fail. Now what?

11.2 Transforming data

11.2.1 Log transform

11.2.2 Arcsine transform

11.2.3 Cube root

11.2.4 Square root transform

11.3 Exercises

11.3.1 Exercise 1

11.3.2 Exercise 2

12

Linear mixed models

Placeholder

12.1 Wilcoxon rank sum test

12.2 Kruskal-Wallis rank sum test

12.2.1 Single factor

12.2.2 Multiple factors

12.3 Generalised linear models

12.3.1 Sign Test

12.3.2 Wilcoxon Signed-Rank Test

12.3.3 Mann-Whitney-Wilcoxon Test

12.3.4 Kruskal-Wallis Test

12.3.5 Generalised linear models (GLM)

12.4 Exercises

12.5 Exercise 1

13

Chi-squared

A chi-squared test is used when one wants to see if there is a relationship between count data of two or more factors.

```
x <- c(A = 20, B = 15, C = 25)
chisq.test(x)
```

```
R>
R> Chi-squared test for given probabilities
R>
R> data:  x
R> X-squared = 2.5, df = 2, p-value = 0.2865
```