

David J. Olive

# Data Science Without Much Math

January 15, 2019





# Preface

This book covers Data Science for students who meet the college entry requirements for Mathematics, but who have not had College Algebra. Hence the student should have had high school algebra and some computer experience. At Southern Illinois University (SIU), the course Math 102 meets the core curriculum. The texts Verzani (2014) and Wilcox (2017) seemed to be close to the appropriate level for such a Data Science course.

A good introductory Statistics text with a College Algebra prerequisite is Moore (2007). This text, and later editions, is used in Math 282 at SIU. This text tries to teach students how to read computer output and to give problems that many Math 282 students would find easy. The *R* software is used. See R Core Team(2016).

Some highlights of this text follow.

- The free software *R* is used.

**Downloading the book's R functions** *dspack.txt* and data files *ds-data.txt* into *R*: The commands

```
source("http://lagrange.math.siu.edu/Olive/dspack.txt")
source("http://lagrange.math.siu.edu/Olive/dsdata.txt")
```

The *R* software is used in this text. See R Core Team (2016).

## Acknowledgements



# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b> .....                           | <b>1</b>  |
|          | 1.1 <b>Introduction</b> .....                       | 1         |
|          | 1.2 <b>The Data Set</b> .....                       | 2         |
|          | 1.3 <b>Summary</b> .....                            | 3         |
|          | 1.4 <b>Complements</b> .....                        | 3         |
|          | 1.5 <b>Problems</b> .....                           | 3         |
| <b>2</b> | <b>Summarizing Data With Graphs</b> .....           | <b>5</b>  |
|          | 2.1 <b>The Bar Graph for Categorical Data</b> ..... | 5         |
|          | 2.2 <b>Graphs for Quantitative Variables</b> .....  | 6         |
|          | 2.3 <b>Summary</b> .....                            | 14        |
|          | 2.4 <b>Complements</b> .....                        | 15        |
|          | 2.5 <b>Problems</b> .....                           | 15        |
| <b>3</b> | <b>Summarizing Data With Statistics</b> .....       | <b>19</b> |
|          | 3.1 <b>Summary</b> .....                            | 19        |
|          | 3.2 <b>Complements</b> .....                        | 19        |
|          | 3.3 <b>Problems</b> .....                           | 19        |
| <b>4</b> | <b>The Normal Distribution</b> .....                | <b>23</b> |
|          | 4.1 <b>Summary</b> .....                            | 23        |
|          | 4.2 <b>Complements</b> .....                        | 24        |
|          | 4.3 <b>Problems</b> .....                           | 24        |
| <b>5</b> | <b>Scatterplots and Correlation</b> .....           | <b>25</b> |
|          | 5.1 <b>Summary</b> .....                            | 25        |
|          | 5.2 <b>Complements</b> .....                        | 25        |
|          | 5.3 <b>Problems</b> .....                           | 25        |
| <b>6</b> | <b>Regression</b> .....                             | <b>27</b> |
|          | 6.1 <b>Summary</b> .....                            | 27        |
|          | 6.2 <b>Complements</b> .....                        | 28        |

|           |  |           |
|-----------|--|-----------|
| 6.3       | Problems   | 28        |
| <b>7</b>  | <b>Sampling</b>                                    | <b>31</b> |
| 7.1       | Nonscientific Surveys                              | 31        |
| 7.2       | Scientific Surveys                                 | 31        |
| 7.3       | Sampling Distribution and the CLT                  | 32        |
| 7.4       | Summary  | 32        |
| 7.5       | Complements  | 33        |
| 7.6       | Problems   | 33        |
| <b>8</b>  | <b>Probability</b>                                 | <b>37</b> |
| 8.1       | Summary  | 37        |
| 8.2       | Complements  | 40        |
| 8.3       | Problems   | 40        |
| <b>9</b>  | <b>Confidence Intervals and Hypothesis Testing</b> | <b>45</b> |
| 9.1       | The $t$ Test and CI                                | 45        |
| 9.2       | Matched Pairs                                      | 45        |
| 9.3       | Two Sample $t$                                     | 45        |
| 9.4       | One Sample $z$ for a Proportion                    | 45        |
| 9.5       | Two Sample $z$ for 2 Proportions                   | 46        |
| 9.6       | Inference for Regression                           | 46        |
| 9.7       | Chi-Squared Tests                                  | 46        |
| 9.8       | Summary  | 46        |
| 9.9       | Complements  | 52        |
| 9.10      | Problems   | 52        |
| <b>10</b> | <b>Classification and Regression Trees</b>         | <b>63</b> |
| 10.1      | Summary  | 63        |
| 10.2      | Complements  | 63        |
| 10.3      | Problems   | 63        |
|           | <b>Index</b>                                       | <b>67</b> |