Fantasy Football Analytics: Statistics, Prediction, and Empiricism Using R

Version 0.0.1

Isaac T. Petersen

To our daughter, Maisie.

Table of contents

Preface **xv**

How to Contribute . . . xv

Open Access . . . xv

License . . . xvi

Citation . . . xvi

About the Author . . . xvii

Accessibility . . . xviii

Acknowledgments . . . xviii

1 Introduction **1**

1.1 About this Book . . . 1

1.2 What is Fantasy Football? . . . 1

1.3 Why Focus on Fantasy Football? . . . 2

1.4 Educational Value . . . 2

1.5 Learning Objectives . . . 3

1.6 Disclosures . . . 4

1.7 Disclaimer . . . 4

2 Intro to Football and Fantasy **5**

2.1 Football . . . 5

2.1.1 The Objective . . . 5

2.1.2 The Roster . . . 5

2.1.3 The Field . . . 8

2.1.4 The Gameplay . . . 10

2.1.5 The Scoring . . . 11

2.1.6 Glossary of Terms 11
2.2 Fantasy Football 13
2.2.1 Overview of Fantasy Football 13
2.2.2 The Fantasy League 14
2.2.3 The Roster of a Fantasy Team 14
2.2.4 Scoring 15

3 Getting Started with R for Data Analysis 19
3.1 Initial Setup 19
3.2 Installing Packages 21
3.3 Load Packages 21
3.4 Using Functions and Arguments 21
3.5 Download Football Data 24
3.5.1 Players 25
3.5.2 Teams 25
3.5.3 Player Info 25
3.5.4 Rosters 25
3.5.5 Game Schedules 25
3.5.6 The Combine 25
3.5.7 Draft Picks 26
3.5.8 Depth Charts 26
3.5.9 Play-By-Play Data 26
3.5.10 4th Down Data 26
3.5.11 Participation 39
3.5.12 Historical Weekly Actual Player Statistics 40
3.5.13 Injuries 40
3.5.14 Snap Counts 40
3.5.15 ESPN QBR 41
3.5.16 NFL Next Gen Stats 41
3.5.17 Advanced Stats from PFR 42
3.5.18 Player Contracts 43

3.5.19 FTN Charting Data 43
3.5.20 Fantasy Player IDs 44
3.5.21 FantasyPros Rankings 44
3.5.22 Expected Fantasy Points 44
3.6 Data Dictionary 45
3.7 Create a Data Frame 45
3.8 Variable Names 46
3.9 Logical Operators 47
3.9.1 Is Equal To: `==` 47
3.9.2 Is Not Equal To: `!=` 47
3.9.3 Is Greater Than: `>` 47
3.9.4 Is Less Than: `<` 47
3.9.5 Is Greater Than or Equal To: `>=` 48
3.9.6 Is Less Than or Equal To: `<=` 48
3.9.7 Is In a Value of Another Vector: `%in%` 48
3.9.8 Is Not In a Value of Another Vector: `!(%in%)` 48
3.9.9 Is Missing: `is.na()` 48
3.9.10 Is Not Missing: `!is.na()` 49
3.9.11 And: `&` 49
3.9.12 Or: `|` 49
3.10 Subset 49
3.10.1 One Variable 50
3.10.2 Particular Rows of One Variable 50
3.10.3 Particular Columns (Variables) 50
3.10.4 Particular Rows 55
3.10.5 Particular Rows and Columns 57
3.11 View Data 58
3.11.1 All Data 58
3.11.2 First 6 Rows/Elements 58
3.12 Data Characteristics 59
3.12.1 Data Structure 59

3.12.2 Data Dimensions 60
3.12.3 Number of Elements 60
3.12.4 Number of Missing Elements 60
3.12.5 Number of Non-Missing Elements 60
3.13 Create New Variables 61
3.14 Recode Variables 61
3.15 Rename Variables 62
3.16 Convert the Types of Variables 62
3.17 Merging/Joins 65
3.17.1 Overview 65
3.17.2 Data Before Merging 66
3.17.3 Types of Joins 67
3.18 Transform Data from Long to Wide 75
3.19 Transform Data from Wide to Long 82
3.20 Loops 83
3.21 Calculations 84
3.21.1 Historical Actual Player Statistics 84
3.21.2 Historical Actual Fantasy Points 87
3.21.3 Player Age 87
3.22 Plotting 92
3.22.1 Rushing Yards per Carry By Player Age 92
3.22.2 Defensive and Offensive EPA per Play 95

4 Player Evaluation 97
4.1 Getting Started 97
4.1.1 Load Packages 97
4.2 Overview 97
4.3 Athletic Profile 98
4.4 Skill 99
4.5 Historical Performance 99
4.5.1 Overview 99

4.5.2 Efficiency 100
4.5.3 Consistency 101
4.6 Health 102
4.7 Age and Career Stage 102
4.8 Situational Factors 103
4.9 Matchups 104
4.10 Cognitive and Motivational Factors 104
4.11 Fantasy Value 105
4.11.1 Sources From Which to Evaluate Fantasy Value 105
4.11.2 Indices to Evaluate Fantasy Value 107
4.12 Putting it Altogether 110

5 The Fantasy Draft 113
5.1 Getting Started 113
5.1.1 Load Packages 113
5.2 Types of Fantasy Drafts 113
5.2.1 Snake Draft 113
5.2.2 Auction Draft 113
5.2.3 Comparison 114
5.3 Draft Strategy 114
5.3.1 Overview 114
5.3.2 Snake Draft 116
5.3.3 Auction Draft 116

6 Research Methods 117
6.1 Getting Started 117
6.1.1 Load Packages 117
6.2 Sample vs Population 117
6.3 Research Designs 118
6.3.1 Experiment 118
6.3.2 Correlational/Observational Study 119

6.3.3 Case Study 121
6.3.4 Other Features of the Research Design 122
6.4 Research Design Validity 124
6.4.1 Internal Validity 124
6.4.2 External Validity 124
6.4.3 Tradeoffs Between Internal and External Validity 125
6.4.4 Conclusion Validity 126
6.5 Mediation vs Moderation 126
6.5.1 Mediation 126
6.5.2 Moderation (i.e., Interaction) 129
6.6 Levels of Measurement 131
6.6.1 Nominal 132
6.6.2 Ordinal 132
6.6.3 Interval 133
6.6.4 Ratio 133
6.7 Psychometrics 134
6.7.1 Measurement Reliability 134
6.7.2 Measurement Validity 136
6.7.3 Reliability vs Validity 138
6.8 Conclusion 139

7 Basic Statistics 141
7.1 Getting Started 141
7.1.1 Load Packages 141
7.2 Descriptive Statistics 141
7.2.1 Center 141
7.2.2 Spread 143
7.2.3 Shape 144
7.2.4 Combination 144
7.3 Scores and Scales 145
7.3.1 Raw Scores 145

7.3.2 z Scores . . . 145
7.4 Inferential Statistics . . . 149
7.4.1 Null Hypothesis Significance Testing . . . 150
7.4.2 Practical Significance . . . 163
7.5 Statistical Decision Tree . . . 166
7.6 Statistical Tests . . . 167
7.6.1 t-Test . . . 167
7.6.2 Analysis of Variance . . . 169
7.6.3 Correlation . . . 170
7.6.4 (Multiple) Regression . . . 170
7.6.5 Chi-Square Test . . . 170
7.6.6 Formulating Statistical Tests in Terms of Partitioned Variance . . . 171
7.6.7 Critical Value . . . 171
7.6.8 Statistical Power . . . 178

8 Correlation Analysis 203
8.1 Getting Started . . . 203
8.1.1 Load Packages . . . 203
8.2 Overview of Correlation . . . 203
8.3 The Correlation Coefficient (r) . . . 203
8.4 Examples . . . 216
8.4.1 Covariance . . . 216
8.4.2 Pearson Correlation . . . 216
8.4.3 Spearman Correlation . . . 216
8.4.4 Nonlinear Correlation . . . 216
8.4.5 Correlation Matrix . . . 216
8.4.6 Correlogram . . . 216
8.5 Correlation Does Not Imply Causation . . . 216
8.6 Conclusion . . . 216
8.7 Session Info . . . 216

9 Multiple Regression **217**

9.1 Getting Started 217

9.1.1 Load Packages 217

9.2 Overview of Multiple Regression 217

9.3 Components 218

9.4 Coefficient of Determination (R^2) 219

9.5 Overfitting 222

9.6 Covariates 224

9.7 Multicollinearity 225

10 Causal Inference **227**

10.1 Getting Started 227

10.1.1 Load Packages 227

10.2 Correlation Does Not Imply Causation 227

10.3 Criteria for Causality 227

10.4 Approaches for Causal Inference 230

10.4.1 Experimental Designs 230

10.4.2 Quasi-Experimental Designs 230

10.5 Causal Diagrams 237

10.5.1 Overview 237

10.5.2 Confounding 245

10.5.3 Mediation 247

10.5.4 Collider Bias 251

10.5.5 Selection Bias 258

10.6 Conclusion 258

11 Heuristics and Cognitive Biases in Prediction **261**

11.1 Getting Started 261

11.1.1 Load Packages 261

11.2 Overview 261

11.3 Examples of Heuristics 263

11.3.1 Availability Heuristic 263
11.3.2 Representativeness Heuristic 263
11.3.3 Anchoring and Adjustment Heuristic 264
11.4 Examples of Cognitive Biases 264
11.4.1 Overconfidence Bias 264
11.4.2 Confirmation Bias 266
11.4.3 Recency Bias 266
11.4.4 Hindsight Bias 267
11.4.5 Loss Aversion Bias 267
11.4.6 Endowment Bias 267
11.4.7 Bandwagon Effect Bias 268
11.4.8 Dunning–Kruger Effect Bias 268
11.5 Examples of Fallacies 268
11.5.1 Base Rate Fallacy 270
11.5.2 Regression Fallacy 270
11.5.3 Hot Hand Fallacy 271
11.5.4 Sunk Cost Fallacy 272
11.5.5 Gambler's Fallacy 272
11.5.6 Conditional Probability Fallacy 275
11.6 Conclusion 275

12 Judgment Versus Actuarial Approaches to Prediction 277
12.1 Getting Started 277
12.1.1 Load Packages 277
12.2 Approaches to Prediction 277
12.2.1 Human Judgment 277
12.2.2 Actuarial/Statistical Method 277
12.2.3 Combining Human Judgment and Statistical Algorithms 278
12.3 Errors in Human Judgment 279
12.4 Humans Versus Computers 281

12.4.1 Advantages of Computers 281
12.4.2 Advantages of Humans 281
12.4.3 Comparison of Evidence 282
12.5 Why Judgment is More Widely Used Than Statistical Formulas 283
12.6 Best Actuarial Approaches to Prediction 284
12.7 Conclusion 285

13 Base Rates 287
13.1 Getting Started 287
13.1.1 Load Packages 287
13.2 Overview 287
13.3 Issues Around Probability 288
13.3.1 Types of Probabilities 288
13.3.2 Confusion of the Inverse 289
13.3.3 Bayes' Theorem 290
13.4 Base Rate of Rookie Performance 294
13.4.1 Quarterbacks 294
13.4.2 Running Backs 294
13.5 How to Account for Base Rates 294
13.5.1 Actuarial Formula 295
13.5.2 Bayesian Updating 295
13.6 Conclusion 301

14 Evaluation of Prediction/Forecasting Accuracy 303
14.1 Getting Started 303
14.1.1 Load Packages 303
14.2 Overview 303
14.3 Types of Accuracy 304
14.3.1 Discrimination 304
14.3.2 Calibration 305
14.3.3 General Accuracy 305

14.4 Prediction of Categorical Outcomes 305
14.5 Prediction of Continuous Outcomes 306
14.6 Threshold-Dependent Accuracy Indices 306
14.6.1 Decision Outcomes 306
14.6.2 Percent Accuracy 307
14.6.3 Percent Accuracy by Chance 308
14.6.4 Predicting from the Base Rate 309
14.6.5 Different Kinds of Errors Have Different Costs 310
14.6.6 Sensitivity, Specificity, PPV, and NPV 311
14.6.7 Signal Detection Theory 321
14.6.8 Accuracy Indices 326
14.7 Threshold-Independent Accuracy Indices 337
14.7.1 General Prediction Accuracy 337
14.7.2 Discrimination 343
14.7.3 Calibration 344
14.8 Integrating the Accuracy Indices 351
14.9 Theory Versus Empiricism 352
14.10 Test Bias 355
14.11 Ways to Improve Prediction Accuracy 355
14.12 Conclusion 357

15 Mythbusters: Putting Fantasy Football Beliefs/Anecdotes to the Test 359

15.1 Getting Started 359
15.1.1 Load Packages 359
15.1.2 Load Data 359
15.2 Do Players Score More Fantasy Points in their Contract Year? 359
15.3 Conclusion 361

References 363

Preface

This is a book in progress—it is incomplete. I will continue to add to and update it as I am able.

How to Contribute

This is an open-access textbook. My goal is to share data analysis strategies for free! Anyone is welcome to contribute to the project. If you would like to contribute, please consider one of the following:

- open an issue[1] or create a pull request[2] on the book's GitHub repository[3].
- buy me a coffee[4]—Support me in developing this (free!) resource for fantasy football analytics… Even a cup of coffee helps me stay awake!

The GitHub repository for the book is located here: https://github.com/isaactpetersen/Fantasy-Football-Analytics-Textbook. If you have data or analysis examples that are you willing to share and include in the book, feel free to contact me.

Open Access

This is an open-access book. This means that it is freely available for anyone to access.

[1] https://github.com/isaactpetersen/Fantasy-Football-Analytics-Textbook/issues
[2] https://github.com/isaactpetersen/Fantasy-Football-Analytics-Textbook/pulls
[3] https://github.com/isaactpetersen/Fantasy-Football-Analytics-Textbook
[4] https://www.buymeacoffee.com/isaactpetersen

License

Figure 1 Creative Commons License

The online version of this book is licensed under the Creative Commons Attribution License[5]. In short, you can use my work as long as you cite it.

Citation

The APA-style citation for the book is:

Petersen, I. T. (2024). *Fantasy football analytics: Statistics, prediction, and empiricism Using R*. Version 0.0.1. University of Iowa Libraries. https://github.com/isaactpetersen/Fantasy-Football-Analytics-Textbook. [INSERT DOI LINK]

The BibTeX citation for the book is:

```
@book{petersenFantasyFootballAnalytics,
 title = {Fantasy football analytics: Statistics, prediction, and empiricism Using R},
  author = {Petersen, Isaac T.},
  year = {2024},
  publisher = {{University of Iowa Libraries}},
  note = {Version 0.0.1},
  doi = {INSERT},
  isbn = {INSERT},
 url = {https://github.com/isaactpetersen/Fantasy-Football-Analytics-Textbook}
}
```

[5] https://creativecommons.org/licenses/by/4.0/

About the Author

I am an Associate Professor in the Department of Psychological and Brain Sciences at the University of Iowa. I am a licensed psychologist with expertise in child clinical psychology. Why am I writing about fantasy football and data analysis? Because fantasy football involves the intersection of two things I love: sports and statistics.

Through my training, I have learned the value of statistics for answering important questions that I find interesting. In graduate training, I came to the realization that statistics are relevant not only for psychology and science, but also for domains that I enjoy as hobbies, including sports and fantasy sports. I have played in a longstanding fantasy football league for over 20 years (since my junior year of high school) with old friends from high school. I wanted to apply what I was learning about statistics to help others improve their performance in fantasy football and to help people—including those who might not otherwise be interested—to learn statistics. So I began blogging online about the value of applying statistics to improve decision making in fantasy football. Apparently, many people were interested in learning statistics when they could apply them to a domain that they find interesting like fantasy football. My blog eventually became FantasyFootballAnalytics.net[6], a website that uses advanced statistics to help people win their fantasy football leagues.

In terms of my `R` and statistics background, I have published many peer-reviewed publications[7] that employ advanced statistical methods, have published a book on psychological assessment[8] (Petersen, 2024b, 2024c) that includes applied examples in `R`, and have published the `petersenlab` `R` package[9] (Petersen, 2024a) on the Comprehensive R Archive Network (CRAN). Several sections in this book come from Petersen (2024c). I am also a co-author of the `ffanalytics` `R` package[10] (Andersen et al., 2024) that provides free utilities for downloading fantasy football projections and additional fantasy-relevant data, and for calculating projected points given your league settings.

[6] http://fantasyfootballanalytics.net
[7] https://developmental-psychopathology.lab.uiowa.edu/publications
[8] https://www.routledge.com/9781032413068
[9] https://cran.r-project.org/web/packages/petersenlab/index.html
[10] https://github.com/FantasyFootballAnalytics/ffanalytics

Accessibility

I strive to follow principles of accessibility[11] (archived at https://perma.cc/8XJ9-Q6QJ) to make the book content accessible to people with visual impairments and physical disabilities. If there are additional ways I can make the content more accessible, please let me know.

Acknowledgments

I thank Dr. Benjamin Motz, who provided consultation and many helpful resources based on his fantasy football statistics class. I also thank key members of FantasyFootballAnalytics.net[12], including Val Pinskiy, Andrew Tungate, Dennis Andersen, and Adam Peterson, who helped develop and provide fantasy football-related resources and who helped sharpen my thinking about the topic. I also thank Professor Patrick Carroll, who taught me the value of statistics for answering important questions.

[11] https://bookdown.org/yihui/rmarkdown-cookbook/html-accessibility.html
[12] http://fantasyfootballanalytics.net

1

Introduction

1.1 About this Book

How can we use information to make predictions about uncertain events? This book is about empiricism (basing theories on observed data) and judgment, prediction, and decision making in the context of uncertainty. The book provides an introduction to modern analytical techniques used to make informed predictions, test theories, and draw conclusions from a given dataset.

This book was originally written for a undergraduate-level course entitled, "Fantasy Football: Predictive Analytics and Empiricism". The chapters provide an overview of topics that each could have its own class and textbook, such as causal inference[1], factor analysis[2], cluster analysis[3], principal component analysis[4], machine learning[5], cognitive biases[6], modern portfolio theory[7], data visualization[8], simulation[9], etc. The book gives readers an overview of the breadth of the approaches to prediction and empiricism. As a consequence, the book does not cover any one technique or approach in great depth.

1.2 What is Fantasy Football?

Fantasy football is an online game where participants assemble (i.e., "draft") imaginary teams composed of real-life National Football League (NFL) players. In this game, participants compete against their opponents (e.g.,

[1] causal-inference.qmd
[2] factor-analysis.qmd
[3] cluster-analysis.qmd
[4] pca.qmd
[5] machine-learning.qmd
[6] cognitive-bias.qmd
[7] modern-portfolio-theory.qmd
[8] data-visualization.qmd
[9] simulation.qmd

friends/coworkers/classmates), accumulating points based on players' actual statistical performances in games. The goal is to outscore one's opponent each week to win matches and ultimately claim victory in the league.

1.3 Why Focus on Fantasy Football?

I was fortunate to have an excellent instructor who taught me the value of learning statistics to answer interesting and important questions. That is, I do not find statistics intrinsically interesting; rather, I find them interesting because of what they allow me to do. Many students find statistics intimidating in part because of how it is typically taught—with examples like dice rolls and coin flips that are (seemingly irrelevant and) boring to students. My contention is that applied examples are a more effective lens to teach many concepts in psychology and data analysis. It can be more engaging and relatable to learn statistics in the applied context of sports, a domain that is more intuitive to many. Many people play fantasy sports. This book involves applying statistics to a particular domain (football). People actually want to learn statistical principles and methods when they can apply them to interesting questions (e.g., sports). In my opinion [and supported by evidence; Motz (2013)], this is a much more effective way of engaging people and teaching statistics than in the context of abstract coin flips and dice rolls. Fantasy football relies heavily on prediction—trying to predict which players will perform best and selecting them accordingly. In this way, fantasy football provides a plethora of decision making opportunities in the face of uncertainty, and a wealth of data for analyzing these decisions. However, unlike many other applied domains in psychology, fantasy football (1) allows a person to see the accuracy of their predictions on a timely basis and (2) provides a safe environment for friendly competition. Thus, it provides a unique domain to evaluate—and improve—the accuracy of various prediction models.

1.4 Educational Value

Skills in statistics, statistical programming, and data analysis are highly valuable. This book includes practical and conceptual tools that build a foundation for critical thinking. The book aims to help readers evaluate theory in the light of evidence (and vice versa) and to refine decision making in the context of uncertainty. Readers will learn about the ways that psychological science (and

related disciplines) poses questions, formulates hypotheses, designs studies to test those questions, and interprets the findings, collectively with the aim to answer questions, improve decision making, and solve problems.

Of course, this is not a traditional psychology textbook. However, the book incorporates important psychological concepts, such as cognitive biases in judgment and prediction, etc. In the modern world of big data, research and society need people who know how to make sense of the information around us. Psychology is in a prime position to teach applied statistics to a wide variety of students, most of whom will not have careers as psychologists. Psychology can teach the importance of statistics given humans' cognitive biases. It can also teach about how these biases can influence how people interpret statistics. This book will teach readers the applications of statistics (prediction) and research methods (empiricism) to answer questions they find interesting, while applying scientific and psychological rigor.

1.5 Learning Objectives

This book aims to help readers accomplish the following learning objectives:

- Apply empirical inference and appreciate the value it provides over speculative supposition.
- Ask educated questions when confronted with decisions in the face of uncertainty.
- Understand human decision making, including common heuristics and cognitive biases and how to mitigate them analytically.
- Engage in critical thinking about causality, including devising plausible alternative explanations for observed effects.
- Understand causal inference including confounding, causal pathways, and counterfactuals.
- Think empirically about human behavior and performance.
- Describe the strengths and weaknesses of humans versus computers in prediction scenarios.
- Apply basic skills in statistical programming using `R` to manipulate and summarize datasets and to conduct data analysis.
- Critically evaluate the strengths and limitations of different statistical models and methodologies used in predicting uncertain events, enhancing their understanding of statistical inference and model selection.
- Use various analytical techniques for predicting the outcome of uncertain events, and for uncovering latent causes of patterns in observed data.
- Interpret findings from various statistical approaches and evaluate the accuracy of predictions.

- Engage in iterative problem-solving processes, refining analytical approaches based on feedback and outcomes, and adapting strategies accordingly.
- Communicate statistical findings and analyses in both written and oral formats, demonstrating proficiency in presenting complex information to diverse audiences.
- Make sense of big data.
- Use practical analytical skills that can be applied in future research and job settings.

1.6 Disclosures

I am the Owner of Fantasy Football Analytics, LLC, which operates https://fantasyfootballanalytics.net.

1.7 Disclaimer

"*This material probably won't win you fantasy football championships. You could take what we learn and apply it to fantasy football and you might become 5 percent more likely to win. Or… Consider the broader relevance of this. You could learn data analysis and figure out ways to apply it to other systems. And you could be making a six-figure salary within the next five years.*" – Benjamin Motz, Ph.D.

2

Intro to Football and Fantasy

This chapter provides a brief primer on (American) football and fantasy football. If you are already familiar with fantasy football, feel free to skip this chapter.

2.1 Football

Football is the most widely watched sport in the United States.[1]

2.1.1 The Objective

The goal in football is for a team to score more points than their opponent. A game lasts 60 minutes, and it is separated into four 15-minute quarters. The team with the most points when the time runs out wins.

2.1.2 The Roster

2.1.2.1 Overview

Each team has 11 players on the field at a time. The particular players who are on the field will depend on the situation, but usually includes one of the three subsets of players:

1. Offense
2. Defense
3. Special Teams

[1] https://news.gallup.com/poll/610046/football-retains-dominant-position-favorite-sport.aspx (archived at https://perma.cc/X2UG-RAAK); https://www.statista.com/statistics/1430289/most-watched-sports-leagues-usa/ (archived at https://perma.cc/JNU6-S96A)

An example formation is depicted in Figure 2.1.

Figure 2.1 An Example Football Formation for the Offense and Defense. The solid line indicates the line of scrimmage. The arrow indicates the direction the offense tries to advance the ball.

2.1.2.2 Offense

The offense is on the field when the team has the ball.

Players on offense include:

- Quarterback (QB)
- Running Back (RB)
  - Halfback (HB) or Tailback (TB)
  - Fullback (FB)
- Wide Receiver (WR)
- Tight End (TE)
- Offensive Linemen (OL), part of the "Offensive Line"
  - Center (C)
  - Offensive Guard (OG)
  - Offensive Tackle (OT)

The quarterback is the most important player on the offense. They help lead the team down the field. The quarterback receives the ball from the Center at the beginning of the play, and they can either hand the ball off (typically to a Running Back or Fullback), pass the ball (typically to a Wide Receiver or

Tight End), or run the ball. Quarterbacks tend to have a strong arm for throwing the ball far and accurately. Some quarterbacks are fast and are considered "dual threats" to pass or run.

Running Backs take a hand-off from the Quarterback to execute a running play (i.e., a rush). They may also catch short passes from the Quarterback or help protect (i.e., block for) the Quarterback from the defensive players who are trying to tackle the Quarterback. Halfbacks and Tailbacks tend to be quick and agile. Fullbacks tend to be strong and powerful.

Wide Receivers catch passes from the Quarterback to execute a passing play. On running plays, they provide protection for the player running the ball (e.g., the Running Back) so the ball carrier can get as far as possible without being tackled. Wide receivers tend to be tall, fast, have good hands (can catch the ball well), and can jump high.

Tight Ends block for running and passing plays, and they catch passes from the Quarterback. Tight ends tend to be strong and have good hands.

Offensive Linemen block for running and passing plays. On passing plays, they provide protection for the Quarterback so the Quarterback has time to pass the ball without being tackled. On running plays, they provide protection for the player running the ball (e.g., the Running Back) so the ball carrier can get as far as possible without being tackled. Offensive Linemen tend to be large so they can provide adequate protection for the Quarterback and Running Back.

2.1.2.3 Defense

The defense is on the field when the team does not have the ball (i.e., when the opposing team has the ball).

Players on defense include:

- Defensive Linemen (DL), part of the "Defensive Line"
  - Defensive End (DE)
  - Defensive Tackle (DT)
- Linebacker (LB)
  - Middle (or Inside) Linebacker (MLB)
  - Outside Linebacker (OLB)
- Defensive Back (DB), part of the "Secondary"
  - Cornerback (CB)
  - Safety (S)
    - Free Safety (FS)
    - Strong Safety (SS)

The players on the defense attempt to tackle the offensive players for as short of gains as possible and attempt to prevent completed passes.

On passing plays, Defensive Linemen try to apply pressure to the Quarterback and try to tackle the Quarterback behind the line of scrimmage before the Quarterback can throw the ball (i.e., a sack). On rushing plays, Defensive Linemen try to tackle the ball carrier to prevent the ball carrier from advancing the ball (i.e., gaining yards). Defensive Linemen tend to be large yet quick so they can apply pressure to the Quarterback.

Linebackers are versatile in that, on a given play, they may attempt to a) "blitz" to sack the Quarterback, b) stop the Running Back, or c) prevent a completed pass. Linebackers tend to be strong yet agile.

Defensive Backs are specialist pass defenders. The main role of Cornerbacks is to cover the Wide Receivers. Safeties serve as the last line of defense for longer passes. Defensive Backs tend to be quick and agile.

2.1.2.4 Special Teams

The special teams involves specialist players who are on the field during all kicking plays including kickoffs, field goals, and punts.

Players on special teams include:

- Kicker (K)
- Punter (P)
- Holder
- Long Snapper
- Punt Returner
- Kick Returner
- and other players intended to block for or to tackle the ball carrier

On a field goal attempt, the Long Snapper snaps the ball to the Holder, who holds the ball for the Kicker. The Kicker attempts field goals and, during kickoffs, kicks the ball to the opposing team. During kickoffs, the Kick Returner catches the kicked ball and returns it for as many yards as possible. During a punt play, the Long Snapper snaps the ball to the Punter who kicks (i.e., punts) the ball to the opposing team. The Punt Returner catches the punted ball and returns it for as many yards as possible.

2.1.3 The Field

The football field is rectangular and is 120 yards long and 53 1/3 yards wide (109.73 m x 48.77 m).[2] At each end of the 120-yard field is a team's end zone.

[2] One yard is equal to three feet. A yard is just smaller than a meter (0.9144 meters).

Each end zone is 10 yards long (9.14 m). Thus, the distance from one end zone to the other end zone is 100 yards (91.44 m). Behind each end zone is a field goal post. A diagram of a football field is depicted in Figure 2.2.

Figure 2.2 A Diagram of a Football Field. The yard markers depict the distance from the nearest end zone. The orange shaded area is called the "red zone", where chances of scoring points are highest. The original figure was modified to depict field goal posts. (Figure retrieved from https://commons.wikimedia.org/wiki/File:American_football_field.svg)

2.1.4 The Gameplay

At the beginning of the game, there is a coin flip to determine which teams receives the ball first and which team takes which side of the field. During the kickoff, the kicking team kicks the ball to the receiving team, who has the option to return the kick. The offense starts their possession at the 25 yard line—if there is no return (i.e., a touchback)—or wherever the kick returner is tackled or goes out of bounds.

The team with the ball (i.e., the offense) has four opportunities ("downs") to advance the ball (i.e., gain) 10 yards. A team can advance the ball either by running it or by throwing (i.e., passing) and catching it. At the end of a rushing play, the ball advances to wherever the ball carrier is tackled or goes out of bounds (i.e., wherever the player is "down"). At the end of a passing play, if the thrown ball is caught (i.e., a completed pass), the ball advances to wherever the ball carrier is tackled or goes out of bounds. If the thrown ball is not caught in bounds before the ball hits the ground (i.e., an incomplete pass), the ball does not advance. Wherever the ball is advanced to dictates where the next play begins. The yard position on the field where the next play takes place from is known as the "line of scrimmage". Neither team can cross the line the line of scrimmage until the next play begins. To begin the play, the ball is placed on the line of scrimmage and the Center gives (or "snaps") the ball to the Quarterback.

If the team advances the ball 10 or more yards within four downs, the team receives a "first down" and is awarded a new set of downs—four more downs to advance the ball 10 more yards. If the team advances the ball all the way to the other team's end zone, they score a touchdown. If the team fails to advance the ball 10 or more yards within four downs, the team loses the ball, and the other team takes possession at that spot on the field. There are risks of giving the other team the ball with a short distance to score. Thus, on fourth down, instead of trying to advance the ball for a first down, a team may choose to kick a field goal—to get points—or to punt.

A field goal involves a kicker kicking the ball with an intent to kick the ball through the field goal posts ("uprights"). To score points by making a field goal, the kicked ball must go between the uprights (extended vertically) and over the cross bar.

Punting involves a punter kicking the ball to the other team with an intent to give their opponent worse field position, thus making it harder for the other team to score. The punting team tries to pin the opponent as close as possible to the opponent's end zone (i.e., as far as possible from the own team's end zone), so they have a longer distance to go to score a touchdown.

There are multiple ways that ball possession can switch from the offense to the other team. After scoring a touchdown, field goal, or safety, there is a kickoff, in which the scoring team kicks the ball to the opponent. Another

way that the ball switches possession to the other team is if the team commits a turnover. The defense can force a turnover by an interception, fumble recovery, or turnover on downs. A turnover due to an interception occurs when a defensive player catches the quarterback's pass. A turnover due to a fumble recovery occurs when an offensive player, who had possession of the ball, loses the ball before being down or scoring a touchdown and the ball is recovered by the opponent. A turnover on downs occurs when the team attempts on fourth down to achieve the remainder of the needed 10 yards to go but fails.

Other football-related situations include tackles for loss and sacks. A tackle for loss occurs when a ball carrier is tackled behind the line of scrimmage. A sack occurs when a Quarterback is tackled with the ball behind the line of scrimmage. A pass defended occurs when a defensive player knocks down the ball in the air so that the indended receiver cannot catch the ball.

2.1.5 The Scoring

The goal of the team with the ball (i.e., the offense) is to score points. It can do this by either advancing the ball into the other team's end zone (6 points) or by kicking a field goal (3 points). Advancing the ball in the other team's end zone is called a touchdown. After a touchdown, the offense chooses to attempt either a point-after-touchdown (PAT) or a two-point conversion. A PAT is a short kick attempt from the 15-yard line (i.e., 15 yards away from the end zone) that, if it goes through the goal posts ("uprights") and over the cross bar, is worth 1 point. A two-point conversion is a single-scoring opportunity from the 3-yard line (i.e., 3 yards away from the end zone). If the offense scores (i.e., advances the ball into the end zone) from the 3-yard line, the team is awarded 2 points.

A team can kick a field goal from any distance as long as the kick goes through the goal posts. The current record for the longest field goal is 66 yards (by Justin Tucker in 2021).

A safety occurs when the offense is tackled with the ball in their own end zone. When a safety occurs, the opposing team (i.e., defense) is awarded two points and the ball.

2.1.6 Glossary of Terms

- running play ("run") or rushing play (or "rush")—the attempt by an offensive player, typically the Running Back or Quarterback, to advance the ball "on the ground" by running it—not by passing it forward
- passing play (or "pass")—the attempt by an offensive player, typically the Quarterback, to advance the ball by throwing it forward to an offensive player

- passing attempt—the attempt to advance the ball by passing it (i.e., a thrown pass)
- rushing attempt—the attempt to advance the ball by running it
- passing completion—a thrown pass that is succesfully caught by an offensive player
- passing incompletion—a thrown pass that is not caught by an offensive player
- passing yards—the distance (in yards) the player advanced the ball by throwing it
- rushing yards—the distance (in yards) the player advanced the ball by running it
- receving yards—the distance (in yards) the player advanced the ball by catching thrown passes and then running with it further upfield
- kick/punt return yards—the distance (in yards) the player advanced the ball by returning kicks or punts
- turnover return yards—the distance (in yards) the player advanced the ball by returning turnovers
- reception—a pass that is caught by the offensive player
- touchdown—advancing the ball into the opponent's end zone either by a) throwing a completed pass that ends up in the end zone, b) running it into the end zone, c) catching it in the end zone, or d) catching it and then running it into the end zone
- passing touchdown—advancing the ball into the opponent's end zone either by throwing a completed pass that ends up in the end zone
- rushing touchdown—advancing the ball into the opponent's end zone either by running it into the end zone
- receiving touchdown—advancing the ball into the opponent's end zone either by catching it in the end zone or by catching it and then running it into the end zone
- kick/punt return touchdown—advancing the ball into the opponent's end zone when returning a kick or punt
- turnover return touchdown—advancing the ball into the opponent's end zone when returning a turnover (i.e., interception or fumble)
- two-point conversion—a single-scoring opportunity from the 3-yard line (i.e., 3 yards away from the end zone) that is an option given to a team that scores a touchdown; if the offense scores (i.e., advances the ball into the end zone) from the 3-yard line, the team is awarded 2 points
- block—when the defense/special teams blocks a kick or field goal by hitting the ball just after it is kicked to prevent the ball from going far
- kickoff—the kicking team kicks the ball to the receiving team, who has the option to return the kick
- field goal—a kicker kicks the ball with an intent to kick the ball through the field goal posts ("uprights"). To score points by making a field goal, the kicked ball must go between the uprights (extended vertically) and over the cross bar. If the field goal attempt is successful, the team gains 3 points.

- point after touchdown (PAT)—a short kick attempt from the 15-yard line (i.e., 15 yards away from the end zone) that, if it goes through the goal posts ("uprights") and over the cross bar, is worth 1 point
- extra point returned—if the defense/special teams returns the ball into the opponent's end zone during a point after touchdown (PAT) attempt, it is worth 2 points
- punt—a punter kicks the ball to the other team with an intent to give their opponent worse field position, thus making it harder for the other team to score
- fumble lost—when an offensive player, who had possession of the ball, loses the ball before being down or scoring a touchdown and the ball is recovered by the opponent
- fumble forced—when a defensive player knocks the ball out of the hands of an offensive player, who had possession of the ball
- fumble recovery—when a defensive player recovers a fumble by the opponent
- interception—when a defensive player catches a pass from an offensive player
- tackle—when a player brings down the ball carrier
- tackle solo—when a player is the main tackler (i.e., the primary player to bring down the ball carrier)
- tackle assist—when a player is one of two or more players who, together, bring down the ball carrier
- tackle for loss—when an offensive player is tackled with the ball behind the line of scrimmage
- sack—when a Quarterback is tackled with the ball behind the line of scrimmage
- pass defended—when a defensive player knocks down the ball in the air so that the indended receiver cannot catch the ball
- safety—when the offense is tackled with the ball in their own end zone

2.2 Fantasy Football

2.2.1 Overview of Fantasy Football

Fantasy football is one of the most widely played games in the history of games. It is estimated that around 62 million people play fantasy sports[3], of whom around 29 million play fantasy football.[4] As noted in the Introduction[5],

[3] https://thefsga.org/industry-demographics/ (archived at https://perma.cc/9PB8-ZDJJ)

[4] https://www.statista.com/topics/10895/fantasy-sports-in-the-us/ (archived at https://perma.cc/8YSN-UUNT)

[5] intro.qmd

fantasy football is an online game where participants assemble (i.e., "draft") imaginary teams composed of real-life National Football League (NFL) players.[6] The participants are in charge of managing and making strategic decisions for their imaginary team to have the best possible team that will score the most points. Thus, the participants are called "managers". Managers make decisions such as selecting which players to draft, selecting which players to play (i.e., "start") on a weekly basis, identifying players to pick up from the remaining pool of available players (i.e., waiver wire), and making trades with other teams. Fantasy football relies heavily on prediction—trying to predict which players will perform best and selecting them accordingly.

2.2.2 The Fantasy League

A fantasy football "league" is composed of various imaginary (i.e., "fantasy") teams—and their associated manager. In the fantasy league, the managers' fantasy teams play against each other. A fantasy league is commonly composed of 8, 10, or 12 fantasy teams, but leagues can have more or fewer teams.

2.2.3 The Roster of a Fantasy Team

On a given roster, a manager has a "starting lineup" and a "bench". Each week, the manager decides which players on their roster to put in the starting lineup, and which to keep on the bench. In many leagues, a starting lineup is composed of offensive players, a kicker, and defense/special teams:

Offensive players:

Table 2.1 Offensive Players in the Starting Lineup

Position	Typical Number of Players in Starting Lineup
Quarterback (QB)	1
Running Back (RB)	2
Wide Receiver (WR)	2
Tight End (TE)	1
Flex Position	1

A "flex position" is a flexible position that can involve a player from various positions: e.g., a Running Back, Wide Receiver, or Tight End.

Kickers:

[6] Fantasy leagues are also available for baseball[7], basketball[8], and many other sports.

- one Kicker (K)

Defense/Special Teams:

- one Team Defense (DST/D/DEF) or multiple Individual Defensive Players (IDP)

2.2.4 Scoring

2.2.4.1 Scoring Overview

In the game of fantasy football, managers accumulate points on a weekly basis based on players' actual statistical performances in NFL games. Managers receive points for only those players who are on their starting lineup (not players on their bench). A manager's goal is to outscore their opponent each week to win matches and ultimately claim victory in the league. Scoring settings can differ from league to league.

Below are common scoring settings for fantasy leagues.

2.2.4.2 Offensive Players

Table 2.2 Common Scoring Settings for Offensive Players

Statistical category	Points
Rushing or receiving TD	6
Returning a kick or punt for a TD	6
Returning or recovering a fumble for a TD	6
Passing TD	4
Passing INT	−2
Fumble lost	−2
Rushing, passing, or receiving 2-point conversion	2
Rushing or receiving yards	1 point per 10 yards
Passing yards	1 point per 25 yards

Note: "TD" = touchdown; "INT" = interception

Other common (but not necessarily standard) statistical categories include:

- receptions (called "point per reception" [PPR] leagues)

- return yards
- passing attempts
- rushing attempts

2.2.4.3 Kickers

Table 2.3 Common Scoring Settings for Kickers

Statistical category	Points
FG made: 50+ yards	5
FG made: 40–49 yards	4
FG made: 39 yards or less	3
Rushing, passing, or receiving 2-point conversion	2
Point after touchdown attempt made	1
Point after touchdown attempt missed	−1
Missed FG: 0–39 yards	−2
Missed FG: 40–49 yards	−1

Note: "FG" = field goal

2.2.4.4 Team Defense/Special Teams

Table 2.4 Common Scoring Settings for Team Defense/Special Teams

Statistical category	Points
Defensive or special teams TD	3
Interception	2
Fumble recovery	2
Blocked punt, PAT, or FG	2
Safety	2
Sack	1

Note: "TD" = touchdown; "PAT" = point after touchdown; "FG" = field goal

2.2.4.5 Individual Defensive Players

Table 2.5 Common Scoring Settings for Individual Defensive Players

Statistical category	Points
Tackle solo	1
Tackle assist	0.5
Tackle for loss	1
Sack	2
Interception	4
Fumble forced	2
Fumble recovery	2
TD	6
Safety	2
Pass defended	1
Blocked kick	2
Extra point returned	2

Note: "TD" = touchdown

Other common (but not necessarily standard) statistical categories include:

- turnover return yards

2.2.4.6 Common Scoring Abbreviations

- "TD" = touchdown
- "INT" = interception
- "yds" = yards
- "ATT" = attempts
- "2-pt conversion" = two-point conversion
- "FG" = field goal
- "PAT" = point after touchdown (i.e., extra point/point after attempt)

3

Getting Started with R for Data Analysis

The book uses R for statistical analyses (http://www.r-project.org). R is a free software environment; you can download it at no charge here: https://cran.r-project.org.

3.1 Initial Setup

To get started, follow the following steps:

1. Install R: https://cran.r-project.org
2. Install RStudio Desktop: https://posit.co/download/rstudio-desktop
3. After installing RStudio, open RStudio and run the following code in the console to install several key R packages:

```
install.packages(
  c("petersenlab","remotes","nflreadr","nflfastR","nfl4th","nflplotR",
  "gsisdecoder","progressr","lubridate","tidyverse","psych"))
```

4. Some necessary packages, including the ffanalytics package, are hosted in GitHub and need to be installed using the following code (after installing the remotes package above):

```
remotes::install_github("FantasyFootballAnalytics/ffanalytics")
```

Note 1: If you are in Dr. Petersen's class

If you are in Dr. Petersen's class, also perform the following steps:

1. Set up a free account on GitHub.com[a].
2. Download GitHub Desktop: https://desktop.github.com
3. Make sure you are logged into your GitHub account on GitHub.com[b].
4. Go to the following GitHub repository: https://github.com/isaactpetersen/QuartoBlogFantasyFootball and complete the following steps:
5. Click "Use this Template" (in the top right of the screen) > "Create a new repository"
6. Make sure the checkbox is selected for the following option: "Include all branches"
7. Make sure your Owner account is selected
8. Specify the repository name to whatever you want, such as `FantasyFootballBlog`
9. Type a brief description, such as `Files for my fantasy football blog`
10. Keep the repository public (this is necessary for generating your blog)
11. Select "Create repository"
12. After creating the new repository, make sure you are on the page of of your new repository and complete the following steps:
13. Click "Settings" (in the topof the screen)
14. Click "Actions" (in the left sidebar) > "General"
15. Make sure the following are selected: - "Read and write permissions" (under "Workflow permissions") - "Allow GitHub Actions to create and approve pull requests" - then click "Save"
16. Click "Pages" (in the left sidebar)
17. Make sure the following are selected: - "Deploy from a branch" (under "Source") - "gh-pages/(root)" (under "Branch") - then click "Save"
18. Clone the repository to your local computer by clicking "Code" > "Open with GitHub Desktop", select the folder where you want the repository to be saved on your local computer, and click "Clone"

[a] https://github.com
[b] https://github.com

3.2 Installing Packages

You can install `R` packages using the following syntax:

```
install.packages("INSERT_PACKAGE_NAME_HERE")
```

For instance, you can use the following code to install the `nflreadr` package:

```
install.packages("nflreadr")
```

3.3 Load Packages

```
library("ffanalytics")
library("nflreadr")
library("nflfastR")
library("nfl4th")
library("nflplotR")
library("progressr")
library("lubridate")
library("tidyverse")
```

3.4 Using Functions and Arguments

You can learn about a particular function and its arguments by entering a question mark before the name of the function:

```
?NAME_OF_FUNCTION()
```

Below, we provide examples for how to learn about and use functions and arguments, by using the `seq()` function as an example. The `seq()` function creates a sequence of numbers. To learn about the `seq()` function, which creates a sequence of numbers, you can execute the following command:

```
?seq()
```

This is what the documentation shows for the `seq()` function in the `Usage` section:

```
seq(
  from = 1,
  to = 1,
  by = ((to - from)/(length.out - 1)),
  length.out = NULL,
  along.with = NULL,
  ...)
```

Based on this information, we know that the `seq()` function takes the following arguments:

- `from`
- `to`
- `by`
- `length.out`
- `along.with`
- `...`

The arguments have default values that are used if the user does not specify values for the arguments. The default values are provided in the `Usage` section and are in Table 3.1:

Table 3.1 Arguments and defaults for the `seq()` function. Arguments with a default of `NULL` are not used unless a value is provided by the user.

Argument	Default Value for Argument
`from`	`1`
`to`	`1`
`by`	`((to - from)/(length.out - 1))`
`length.out`	`NULL`
`along.with`	`NULL`

What each argument represents (i.e., the meaning of `from`, `to`, `by`, etc.) is provided in the `Arguments` section of the documentation. You can specify a function and its arguments either by providing values for each argument in the order indicated by the function, or by naming its arguments.

Here is an example of providing values to the arguments in the order indicated by the function, to create a sequence of numbers from 1 to 9:

```
seq(1, 9)
```

```
[1] 1 2 3 4 5 6 7 8 9
```

Here is an example of providing values to the arguments by naming its arguments:

```
seq(
  from = 1,
  to = 9,
  by = 1)
```

```
[1] 1 2 3 4 5 6 7 8 9
```

If you provide values to arguments by naming the arguments, you can reorder the arguments and get the same answer:

```
seq(
  by = 1,
  to = 9,
  from = 1)
```

```
[1] 1 2 3 4 5 6 7 8 9
```

There are various combinations of arguments that one could use to obtain the same result. For instance, here is code to generate a sequence from 1 to 9 by 2:

```
seq(
  from = 1,
  to = 9,
  by = 2)
```

```
[1] 1 3 5 7 9
```

Or, alternatively, you could specify the length of the desired sequence (5 values):

```
seq(
  from = 1,
  to = 9,
  length.out = 5)
```

```
[1] 1 3 5 7 9
```

If you want to generate a series with decimal values, you could specify a long desired sequence of 81 values:

```r
seq(
  from = 1,
  to = 9,
  length.out = 81)
```

```
 [1] 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8
[20] 2.9 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2 4.3 4.4 4.5 4.6 4.7
[39] 4.8 4.9 5.0 5.1 5.2 5.3 5.4 5.5 5.6 5.7 5.8 5.9 6.0 6.1 6.2 6.3 6.4 6.5 6.6
[58] 6.7 6.8 6.9 7.0 7.1 7.2 7.3 7.4 7.5 7.6 7.7 7.8 7.9 8.0 8.1 8.2 8.3 8.4 8.5
[77] 8.6 8.7 8.8 8.9 9.0
```

This is equivalent to specifying a sequence from 1 to 9 by 0.1:

```r
seq(
  from = 1,
  to = 9,
  by = 0.1)
```

```
 [1] 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8
[20] 2.9 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2 4.3 4.4 4.5 4.6 4.7
[39] 4.8 4.9 5.0 5.1 5.2 5.3 5.4 5.5 5.6 5.7 5.8 5.9 6.0 6.1 6.2 6.3 6.4 6.5 6.6
[58] 6.7 6.8 6.9 7.0 7.1 7.2 7.3 7.4 7.5 7.6 7.7 7.8 7.9 8.0 8.1 8.2 8.3 8.4 8.5
[77] 8.6 8.7 8.8 8.9 9.0
```

Hopefully, that provides an example for how to learn about a particular function, its arguments, and how to use them.

3.5 Download Football Data

Below, we provide examples for how to download various types of National Football League (NFL) data. For additional resources, Congelio (2023) provides a helpful introductory text for working with NFL data in R.

3.5.1 Players

```
nfl_players <- progressr::with_progress(
  nflreadr::load_players())
```

3.5.2 Teams

```
nfl_teams <- progressr::with_progress(
  nflreadr::load_teams(current = TRUE))
```

3.5.3 Player Info

3.5.4 Rosters

A Data Dictionary for rosters is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_rosters.html

```
nfl_rosters <- progressr::with_progress(
  nflreadr::load_rosters(seasons = TRUE))

nfl_rosters_weekly <- progressr::with_progress(
  nflreadr::load_rosters_weekly(seasons = TRUE))
```

3.5.5 Game Schedules

A Data Dictionary for game schedules data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_schedules.html

```
nfl_schedules <- progressr::with_progress(
  nflreadr::load_schedules(seasons = TRUE))
```

3.5.6 The Combine

A Data Dictionary for data from the combine is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_combine.html

```
nfl_combine <- progressr::with_progress(
  nflreadr::load_combine(seasons = TRUE))
```

3.5.7 Draft Picks

A Data Dictionary for draft picks data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_draft_picks.html

```
nfl_draftPicks <- progressr::with_progress(
  nflreadr::load_draft_picks(seasons = TRUE))
```

3.5.8 Depth Charts

A Data Dictionary for data from weekly depth charts is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_depth_charts.html

```
nfl_depthCharts <- progressr::with_progress(
  nflreadr::load_depth_charts(seasons = TRUE))
```

3.5.9 Play-By-Play Data

To download play-by-play data from prior weeks and seasons, we can use the `load_pbp()` function of the `nflreadr` package. We add a progress bar using the `with_progress()` function from the `progressr` package because it takes a while to run. A Data Dictionary for the play-by-play data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_pbp.html

> **Note 2: Downloading play-by-play data**
>
> Note: the following code takes a while to run.

```
nfl_pbp <- progressr::with_progress(
  nflreadr::load_pbp(seasons = TRUE))
```

3.5.10 4th Down Data

```
nfl_4thdown <- nfl4th::load_4th_pbp(seasons = 2014:2023)
```

```
[04:20:39] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:04] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:05] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:05] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:05] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:21:05] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:11] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:15] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:16] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:23] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:21:30] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:50] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:21:56] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:22:00] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:22:01] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:22:08] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:22:15] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:22:36] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:22:42] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:22:45] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:22:46] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:22:53] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:23:00] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:23:21] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:23:27] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:23:31] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:23:31] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:23:38] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:23:45] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:24:04] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:24:10] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:24:14] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:24:14] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:24:20] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:24:27] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:24:47] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:24:53] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:24:57] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:24:57] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:25:04] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:25:11] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:25:29] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:25:35] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:25:38] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:25:39] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:25:44] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:25:51] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:26:12] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:26:18] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:26:22] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:26:22] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:26:28] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:26:36] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:26:57] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:27:03] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:27:07] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:27:07] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:27:14] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:27:22] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:27:43] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

```
[04:27:50] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:27:54] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:27:54] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.

[04:28:02] WARNING: src/learner.cc:553:
 If you are loading a serialized model (like pickle in Python, RDS in R) generated by
 older XGBoost, please export the model by calling `Booster.save_model` from that version
  first, then load it back in current version. See:

   https://xgboost.readthedocs.io/en/latest/tutorials/saving_model.html

 for more details about differences between saving model and serializing.
```

3.5.11 Participation

A Data Dictionary for the participation data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_participation.html

```r
nfl_participation <- progressr::with_progress(
  nflreadr::load_participation(
    seasons = TRUE,
```

```
    include_pbp = TRUE))
```

3.5.12 Historical Weekly Actual Player Statistics

We can download historical week-by-week actual player statistics using the `load_player_stats()` function from the `nflreadr` package. A Data Dictionary for statistics for offensive players is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_player_stats.html. A Data Dictionary for statistics for defensive players is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_player_stats_def.html.

```
nfl_actualStats_offense_weekly <- progressr::with_progress(
  nflreadr::load_player_stats(
    seasons = TRUE,
    stat_type = "offense"))

nfl_actualStats_defense_weekly <- progressr::with_progress(
  nflreadr::load_player_stats(
    seasons = TRUE,
    stat_type = "defense"))

nfl_actualStats_kicking_weekly <- progressr::with_progress(
  nflreadr::load_player_stats(
    seasons = TRUE,
    stat_type = "kicking"))
```

3.5.13 Injuries

A Data Dictionary for injury data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_injuries.html

```
nfl_injuries <- progressr::with_progress(
  nflreadr::load_injuries(seasons = TRUE))
```

3.5.14 Snap Counts

A Data Dictionary for snap counts data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_snap_counts.html

```
nfl_snapCounts <- progressr::with_progress(
  nflreadr::load_snap_counts(seasons = TRUE))
```

3.5.15 ESPN QBR

A Data Dictionary for ESPN QBR data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_espn_qbr.html

```
nfl_espnQBR_seasonal <- progressr::with_progress(
  nflreadr::load_espn_qbr(
    seasons = TRUE,
    summary_type = c("season")))

nfl_espnQBR_weekly <- progressr::with_progress(
  nflreadr::load_espn_qbr(
    seasons = TRUE,
    summary_type = c("weekly")))

nfl_espnQBR_weekly$game_week <- as.character(nfl_espnQBR_weekly$game_week)

nfl_espnQBR <- bind_rows(
  nfl_espnQBR_seasonal,
  nfl_espnQBR_weekly
)
```

3.5.16 NFL Next Gen Stats

A Data Dictionary for NFL Next Gen Stats data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_nextgen_stats.html

```
nfl_nextGenStats_pass_weekly <- progressr::with_progress(
  nflreadr::load_nextgen_stats(
    seasons = TRUE,
    stat_type = c("passing")))

nfl_nextGenStats_rush_weekly <- progressr::with_progress(
  nflreadr::load_nextgen_stats(
    seasons = TRUE,
    stat_type = c("rushing")))

nfl_nextGenStats_rec_weekly <- progressr::with_progress(
  nflreadr::load_nextgen_stats(
```

```
    seasons = TRUE,
    stat_type = c("receiving")))

nfl_nextGenStats_weekly <- bind_rows(
  nfl_nextGenStats_pass_weekly,
  nfl_nextGenStats_rush_weekly,
  nfl_nextGenStats_rec_weekly
)
```

3.5.17 Advanced Stats from PFR

A Data Dictionary for PFR passing data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_pfr_passing.html

```
nfl_advancedStatsPFR_pass_seasonal <- progressr::with_progress(
  nflreadr::load_pfr_advstats(
    seasons = TRUE,
    stat_type = c("pass"),
    summary_level = c("season")))

nfl_advancedStatsPFR_pass_weekly <- progressr::with_progress(
  nflreadr::load_pfr_advstats(
    seasons = TRUE,
    stat_type = c("pass"),
    summary_level = c("week")))

nfl_advancedStatsPFR_rush_seasonal <- progressr::with_progress(
  nflreadr::load_pfr_advstats(
    seasons = TRUE,
    stat_type = c("rush"),
    summary_level = c("season")))

nfl_advancedStatsPFR_rush_weekly <- progressr::with_progress(
  nflreadr::load_pfr_advstats(
    seasons = TRUE,
    stat_type = c("rush"),
    summary_level = c("week")))

nfl_advancedStatsPFR_rec_seasonal <- progressr::with_progress(
  nflreadr::load_pfr_advstats(
    seasons = TRUE,
    stat_type = c("rec"),
    summary_level = c("season")))
```

```
nfl_advancedStatsPFR_rec_weekly <- progressr::with_progress(
  nflreadr::load_pfr_advstats(
    seasons = TRUE,
    stat_type = c("rec"),
    summary_level = c("week")))

nfl_advancedStatsPFR_def_seasonal <- progressr::with_progress(
  nflreadr::load_pfr_advstats(
    seasons = TRUE,
    stat_type = c("def"),
    summary_level = c("season")))

nfl_advancedStatsPFR_def_weekly <- progressr::with_progress(
  nflreadr::load_pfr_advstats(
    seasons = TRUE,
    stat_type = c("def"),
    summary_level = c("week")))

nfl_advancedStatsPFR <- bind_rows(
  nfl_advancedStatsPFR_pass_seasonal,
  nfl_advancedStatsPFR_pass_weekly,
  nfl_advancedStatsPFR_rush_seasonal,
  nfl_advancedStatsPFR_rush_weekly,
  nfl_advancedStatsPFR_rec_seasonal,
  nfl_advancedStatsPFR_rec_weekly,
  nfl_advancedStatsPFR_def_seasonal,
  nfl_advancedStatsPFR_def_weekly,
)
```

3.5.18 Player Contracts

A Data Dictionary for player contracts data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_contracts.html

```
nfl_playerContracts <- progressr::with_progress(
  nflreadr::load_contracts())
```

3.5.19 FTN Charting Data

A Data Dictionary for FTN Charting data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_ftn_charting.html

```
nfl_ftnCharting <- progressr::with_progress(
  nflreadr::load_ftn_charting(seasons = TRUE))
```

3.5.20 Fantasy Player IDs

A Data Dictionary for fantasy player ID data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_ff_playerids.html

```
nfl_playerIDs <- progressr::with_progress(
  nflreadr::load_ff_playerids())
```

3.5.21 FantasyPros Rankings

A Data Dictionary for FantasyPros ranking data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_ff_rankings.html

```
#nfl_rankings <- progressr::with_progress( # currently throws error
#  nflreadr::load_ff_rankings(type = "all"))

nfl_rankings_draft <- progressr::with_progress(
  nflreadr::load_ff_rankings(type = "draft"))

nfl_rankings_weekly <- progressr::with_progress(
  nflreadr::load_ff_rankings(type = "week"))

nfl_rankings <- bind_rows(
  nfl_rankings_draft,
  nfl_rankings_weekly
)
```

3.5.22 Expected Fantasy Points

A Data Dictionary for expected fantasy points data is located at the following link: https://nflreadr.nflverse.com/articles/dictionary_ff_opportunity.html

```
nfl_expectedFantasyPoints_weekly <- progressr::with_progress(
  nflreadr::load_ff_opportunity(
    seasons = TRUE,
    stat_type = "weekly",
    model_version = "latest"
  ))
```

```
nfl_expectedFantasyPoints_pass <- progressr::with_progress(
  nflreadr::load_ff_opportunity(
    seasons = TRUE,
    stat_type = "pbp_pass",
    model_version = "latest"
  ))

nfl_expectedFantasyPoints_rush <- progressr::with_progress(
  nflreadr::load_ff_opportunity(
    seasons = TRUE,
    stat_type = "pbp_rush",
    model_version = "latest"
  ))

nfl_expectedFantasyPoints_weekly$season <- as.integer(nfl_expectedFantasyPoints_weekly$season)

nfl_expectedFantasyPoints_offense <- bind_rows(
  nfl_expectedFantasyPoints_pass,
  nfl_expectedFantasyPoints_rush
)
```

3.6 Data Dictionary

Data Dictionaries are metadata that describe the meaning of the variables in a datset. You can find Data Dictionaries for the various NFL datasets at the following link: https://nflreadr.nflverse.com/articles/index.html.

3.7 Create a Data Frame

Here is an example of creating a data frame:

```
players <- data.frame(
  ID = 1:12,
  name = c(
    "Ken Cussion",
    "Ben Sacked",
```

```
    "Chuck Downfield",
    "Ron Ingback",
    "Rhonda Ball",
    "Hugo Long",
    "Lionel Scrimmage",
    "Drew Blood",
    "Chase Emdown",
    "Justin Time",
    "Spike D'Ball",
    "Isac Ulooz"),
  position = c("QB","QB","QB","RB","RB","WR","WR","WR","WR","TE","TE","LB"),
  age = c(40, 30, 24, 20, 18, 23, 27, 32, 26, 23, NA, 37)
  )

fantasyPoints <- data.frame(
  ID = c(2, 7, 13, 14),
  fantasyPoints = c(250, 170, 65, 15)
)
```

3.8 Variable Names

To see the names of variables in a data frame, use the following syntax:

```
names(nfl_players)
```

```
 [1] "status"                   "display_name"
 [3] "first_name"               "last_name"
 [5] "esb_id"                   "gsis_id"
 [7] "suffix"                   "birth_date"
 [9] "college_name"             "position_group"
[11] "position"                 "jersey_number"
[13] "height"                   "weight"
[15] "years_of_experience"      "team_abbr"
[17] "team_seq"                 "current_team_id"
[19] "football_name"            "entry_year"
[21] "rookie_year"              "draft_club"
[23] "college_conference"       "status_description_abbr"
[25] "status_short_description" "gsis_it_id"
[27] "short_name"               "smart_id"
[29] "headshot"                 "draft_number"
```

```
[31] "uniform_number"           "draft_round"
[33] "season"
```

```r
names(players)
```

```
[1] "ID"       "name"     "position" "age"
```

```r
names(fantasyPoints)
```

```
[1] "ID"            "fantasyPoints"
```

3.9 Logical Operators

3.9.1 Is Equal To: ==

```r
players$position == "RB"
```

```
 [1] FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

3.9.2 Is Not Equal To: !=

```r
players$position != "RB"
```

```
 [1]  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

3.9.3 Is Greater Than: >

```r
players$age > 30
```

```
 [1]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE   NA  TRUE
```

3.9.4 Is Less Than: <

```
players$age < 30
```

```
 [1] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE    NA FALSE
```

3.9.5 Is Greater Than or Equal To: `>=`

```
players$age >= 30
```

```
 [1]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE    NA  TRUE
```

3.9.6 Is Less Than or Equal To: `<=`

```
players$age <= 30
```

```
 [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE    NA FALSE
```

3.9.7 Is In a Value of Another Vector: `%in%`

```
players$position %in% c("RB","WR")
```

```
 [1] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
```

3.9.8 Is Not In a Value of Another Vector: `!(%in%)`

```
!(players$position %in% c("RB","WR"))
```

```
 [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
```

3.9.9 Is Missing: `is.na()`

```
is.na(players$age)
```

```
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
```

3.9.10 Is Not Missing: `!is.na()`

```
!is.na(players$age)
```

```
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
```

3.9.11 And: `&`

```
players$position == "WR" & players$age > 26
```

```
 [1] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
```

3.9.12 Or: `|`

```
players$position == "WR" | players$age > 23
```

```
 [1]  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE   NA  TRUE
```

3.10 Subset

To subset a data frame, use brackets to specify the subset of rows and columns to keep, where the value/vector before the comma specifies the rows to keep, and the value/vector after the comma specifies the columns to keep:

```
dataframe[rowsToKeep, columnsToKeep]
```

You can subset by using any of the following:

- numeric indices of the rows/columns to keep (or drop)
- names of the rows/columns to keep (or drop)
- values of `TRUE` and `FALSE` corresponding to which rows/columns to keep

3.10.1 One Variable

To subset one variable, use the following syntax:

```
players$name
```

```
 [1] "Ken Cussion"     "Ben Sacked"      "Chuck Downfield"  "Ron Ingback"
 [5] "Rhonda Ball"     "Hugo Long"       "Lionel Scrimmage" "Drew Blood"
 [9] "Chase Emdown"    "Justin Time"     "Spike D'Ball"     "Isac Ulooz"
```

or:

```
players[,"name"]
```

```
 [1] "Ken Cussion"     "Ben Sacked"      "Chuck Downfield"  "Ron Ingback"
 [5] "Rhonda Ball"     "Hugo Long"       "Lionel Scrimmage" "Drew Blood"
 [9] "Chase Emdown"    "Justin Time"     "Spike D'Ball"     "Isac Ulooz"
```

3.10.2 Particular Rows of One Variable

To subset one variable, use the following syntax:

```
players$name[which(players$position == "RB")]
```

```
[1] "Ron Ingback" "Rhonda Ball"
```

or:

```
players[which(players$position == "RB"), "name"]
```

```
[1] "Ron Ingback" "Rhonda Ball"
```

3.10.3 Particular Columns (Variables)

To subset particular columns/variables, use the following syntax:

3.10.3.1 Base R

```r
subsetVars <- c("name","age")

players[,c(2,4)]
```

```
               name age
1       Ken Cussion  40
2        Ben Sacked  30
3   Chuck Downfield  24
4       Ron Ingback  20
5       Rhonda Ball  18
6         Hugo Long  23
7  Lionel Scrimmage  27
8        Drew Blood  32
9      Chase Emdown  26
10      Justin Time  23
11     Spike D'Ball  NA
12       Isac Ulooz  37
```

```r
players[,c("name","age")]
```

```
               name age
1       Ken Cussion  40
2        Ben Sacked  30
3   Chuck Downfield  24
4       Ron Ingback  20
5       Rhonda Ball  18
6         Hugo Long  23
7  Lionel Scrimmage  27
8        Drew Blood  32
9      Chase Emdown  26
10      Justin Time  23
11     Spike D'Ball  NA
12       Isac Ulooz  37
```

```r
players[,subsetVars]
```

```
               name age
1       Ken Cussion  40
2        Ben Sacked  30
3   Chuck Downfield  24
4       Ron Ingback  20
5       Rhonda Ball  18
```

```
6         Hugo Long  23
7  Lionel Scrimmage  27
8        Drew Blood  32
9      Chase Emdown  26
10      Justin Time  23
11     Spike D'Ball  NA
12       Isac Ulooz  37
```

Or, to drop columns:

```r
dropVars <- c("name","age")

players[,-c(2,4)]
```

```
   ID position
1   1       QB
2   2       QB
3   3       QB
4   4       RB
5   5       RB
6   6       WR
7   7       WR
8   8       WR
9   9       WR
10 10       TE
11 11       TE
12 12       LB
```

```r
players[,!(names(players) %in% c("name","age"))]
```

```
   ID position
1   1       QB
2   2       QB
3   3       QB
4   4       RB
5   5       RB
6   6       WR
7   7       WR
8   8       WR
9   9       WR
10 10       TE
11 11       TE
12 12       LB
```

```
players[,!(names(players) %in% dropVars)]
```

```
   ID position
1   1       QB
2   2       QB
3   3       QB
4   4       RB
5   5       RB
6   6       WR
7   7       WR
8   8       WR
9   9       WR
10 10       TE
11 11       TE
12 12       LB
```

3.10.3.2 Tidyverse

```
players %>%
  select(name, age)
```

```
               name age
1       Ken Cussion  40
2        Ben Sacked  30
3   Chuck Downfield  24
4       Ron Ingback  20
5       Rhonda Ball  18
6         Hugo Long  23
7  Lionel Scrimmage  27
8        Drew Blood  32
9      Chase Emdown  26
10      Justin Time  23
11     Spike D'Ball  NA
12       Isac Ulooz  37
```

```
players %>%
  select(name:age)
```

```
               name position age
1       Ken Cussion       QB  40
2        Ben Sacked       QB  30
3   Chuck Downfield       QB  24
```

```
4       Ron Ingback       RB  20
5       Rhonda Ball       RB  18
6         Hugo Long       WR  23
7  Lionel Scrimmage       WR  27
8        Drew Blood       WR  32
9      Chase Emdown       WR  26
10      Justin Time       TE  23
11     Spike D'Ball       TE  NA
12       Isac Ulooz       LB  37
```

```
players %>%
  select(all_of(subsetVars))
```

```
               name age
1       Ken Cussion  40
2        Ben Sacked  30
3   Chuck Downfield  24
4       Ron Ingback  20
5       Rhonda Ball  18
6         Hugo Long  23
7  Lionel Scrimmage  27
8        Drew Blood  32
9      Chase Emdown  26
10      Justin Time  23
11     Spike D'Ball  NA
12       Isac Ulooz  37
```

Or, to drop columns:

```
players %>%
  select(-name, -age)
```

```
   ID position
1   1       QB
2   2       QB
3   3       QB
4   4       RB
5   5       RB
6   6       WR
7   7       WR
8   8       WR
9   9       WR
10 10       TE
11 11       TE
12 12       LB
```

```r
players %>%
  select(-c(name:age))
```

```
   ID
1   1
2   2
3   3
4   4
5   5
6   6
7   7
8   8
9   9
10 10
11 11
12 12
```

```r
players %>%
  select(-all_of(dropVars))
```

```
   ID position
1   1       QB
2   2       QB
3   3       QB
4   4       RB
5   5       RB
6   6       WR
7   7       WR
8   8       WR
9   9       WR
10 10       TE
11 11       TE
12 12       LB
```

3.10.4 Particular Rows

To subset particular rows, use the following syntax:

3.10.4.1 Base R

```
subsetRows <- c(4,5)

players[c(4,5),]
```

```
  ID        name position age
4  4 Ron Ingback       RB  20
5  5 Rhonda Ball       RB  18
```

```
players[subsetRows,]
```

```
  ID        name position age
4  4 Ron Ingback       RB  20
5  5 Rhonda Ball       RB  18
```

```
players[which(players$position == "RB"),]
```

```
  ID        name position age
4  4 Ron Ingback       RB  20
5  5 Rhonda Ball       RB  18
```

3.10.4.2 Tidyverse

```
players %>%
  filter(position == "WR")
```

```
  ID             name position age
1  6        Hugo Long       WR  23
2  7 Lionel Scrimmage       WR  27
3  8       Drew Blood       WR  32
4  9     Chase Emdown       WR  26
```

```
players %>%
  filter(position == "WR", age <= 26)
```

```
  ID         name position age
1  6    Hugo Long       WR  23
2  9 Chase Emdown       WR  26
```

```
players %>%
  filter(position == "WR" | age >= 26)
```

```
  ID             name position age
1  1      Ken Cussion       QB  40
2  2       Ben Sacked       QB  30
3  6        Hugo Long       WR  23
4  7 Lionel Scrimmage       WR  27
5  8       Drew Blood       WR  32
6  9     Chase Emdown       WR  26
7 12       Isac Ulooz       LB  37
```

3.10.5 Particular Rows and Columns

To subset particular rows and columns, use the following syntax:

3.10.5.1 Base R

```
players[c(4,5), c(2,4)]
```

```
         name age
4 Ron Ingback  20
5 Rhonda Ball  18
```

```
players[subsetRows, subsetVars]
```

```
         name age
4 Ron Ingback  20
5 Rhonda Ball  18
```

```
players[which(players$position == "RB"), subsetVars]
```

```
         name age
4 Ron Ingback  20
5 Rhonda Ball  18
```

3.10.5.2 Tidyverse

```
players %>%
  filter(position == "RB") %>%
  select(all_of(subsetVars))
```

```
         name age
1 Ron Ingback  20
2 Rhonda Ball  18
```

3.11 View Data

3.11.1 All Data

To view data, use the following syntax:

```
View(players)
```

3.11.2 First 6 Rows/Elements

To view only the first six rows (if a data frame) or elements (if a vector), use the following syntax:

```
head(nfl_players)
```

```
# A tibble: 6 x 33
  status display_name    first_name last_name esb_id  gsis_id suffix birth_date
  <chr>  <chr>           <chr>      <chr>     <chr>   <chr>   <chr>  <chr>
1 RET    'Omar Ellison    'Omar      Ellison   ELL711~ 00-000~ <NA>   <NA>
2 ACT    A'Shawn Robinson A'Shawn    Robinson  ROB367~ 00-003~ <NA>   1995-03-21
3 ACT    A.J. Arcuri      A.J.       Arcuri    ARC716~ 00-003~ <NA>   <NA>
4 RES    A.J. Bouye       Arlandus   Bouye     BOU651~ 00-003~ <NA>   1991-08-16
5 ACT    A.J. Brown       Arthur     Brown     BRO413~ 00-003~ <NA>   1997-06-30
6 ACT    A.J. Cann        Aaron      Cann      CAN364~ 00-003~ <NA>   1991-10-03
# i 25 more variables: college_name <chr>, position_group <chr>,
#   position <chr>, jersey_number <int>, height <dbl>, weight <int>,
#   years_of_experience <chr>, team_abbr <chr>, team_seq <int>,
#   current_team_id <chr>, football_name <chr>, entry_year <int>,
#   rookie_year <int>, draft_club <chr>, college_conference <chr>,
#   status_description_abbr <chr>, status_short_description <chr>,
#  gsis_it_id <int>, short_name <chr>, smart_id <chr>, headshot <chr>, ...
```

```r
head(nfl_players$display_name)
```

```
[1] "'Omar Ellison"   "A'Shawn Robinson" "A.J. Arcuri"     "A.J. Bouye"
[5] "A.J. Brown"       "A.J. Cann"
```

3.12 Data Characteristics

3.12.1 Data Structure

```r
str(nfl_players)
```

```
nflvrs_d [20,039 x 33] (S3: nflverse_data/tbl_df/tbl/data.table/data.frame)
 $ status                  : chr [1:20039] "RET" "ACT" "ACT" "RES" ...
 $ display_name        : chr [1:20039] "'Omar Ellison" "A'Shawn Robinson" "A.J. Arcuri" "A.J. Bouye" ..
 $ first_name         : chr [1:20039] "'Omar" "A'Shawn" "A.J." "Arlandus" ...
 $ last_name          : chr [1:20039] "Ellison" "Robinson" "Arcuri" "Bouye" ...
 $ esb_id            : chr [1:20039] "ELL711319" "ROB367960" "ARC716900" "BOU651714" ...
 $ gsis_id           : chr [1:20039] "00-0004866" "00-0032889" "00-0037845" "00-0030228" ...
 $ suffix                  : chr [1:20039] NA NA NA NA ...
 $ birth_date         : chr [1:20039] NA "1995-03-21" NA "1991-08-16" ...
 $ college_name        : chr [1:20039] NA "Alabama" "Michigan State" "Central Florida" ...
 $ position_group          : chr [1:20039] "WR" "DL" "OL" "DB" ...
 $ position                : chr [1:20039] "WR" "DT" "T" "CB" ...
 $ jersey_number         : int [1:20039] 84 91 61 24 11 60 6 81 63 20 ...
 $ height               : num [1:20039] 73 76 79 72 72 75 76 69 76 72 ...
 $ weight            : int [1:20039] 200 330 320 191 226 325 220 190 280 183 ...
 $ years_of_experience     : chr [1:20039] "2" "8" "2" "8" ...
 $ team_abbr               : chr [1:20039] "LAC" "NYG" "LA" "CAR" ...
 $ team_seq                : int [1:20039] NA 1 NA 1 1 1 1 NA NA NA ...
 $ current_team_id       : chr [1:20039] "4400" "3410" "2510" "0750" ...
 $ football_name          : chr [1:20039] NA "A'Shawn" "A.J." "A.J." ...
 $ entry_year         : int [1:20039] NA 2016 2022 2013 2019 2015 2019 NA NA NA ...
 $ rookie_year         : int [1:20039] NA 2016 2022 2013 2019 2015 2019 NA NA NA ...
 $ draft_club              : chr [1:20039] NA "DET" "LA" NA ...
 $ college_conference    : chr [1:20039] NA "Southeastern Conference" "Big Ten Conference" "American A
 $ status_description_abbr : chr [1:20039] NA "A01" "A01" "R01" ...
 $ status_short_description: chr [1:20039] NA "Active" "Active" "R/Injured" ...
 $ gsis_it_id         : int [1:20039] NA 43335 54726 40688 47834 42410 48335 NA NA NA ...
 $ short_name         : chr [1:20039] NA "A.Robinson" "A.Arcuri" "A.Bouye" ...
```

```
 $ smart_id           : chr [1:20039] "3200454c-4c71-1319-728e-d49d3d236f8f" "3200524f-4236-7960-bf20
 $ headshot           : chr [1:20039] NA "https://static.www.nfl.com/image/private/f_auto,q_auto/leag
 $ draft_number         : int [1:20039] NA 46 261 NA 51 67 NA NA NA NA ...
 $ uniform_number           : chr [1:20039] NA "91" "61" "24" ...
 $ draft_round              : chr [1:20039] NA NA NA NA ...
 $ season               : int [1:20039] NA NA NA NA NA NA NA NA NA NA ...
 - attr(*, "nflverse_type")= chr "players"
 - attr(*, "nflverse_timestamp")= POSIXct[1:1], format: "2024-03-01 01:18:40"
```

3.12.2 Data Dimensions

Number of rows and columns:

```r
dim(nfl_players)
```

```
[1] 20039    33
```

3.12.3 Number of Elements

```r
length(nfl_players$display_name)
```

```
[1] 20039
```

3.12.4 Number of Missing Elements

```r
length(nfl_players$college_name[which(is.na(nfl_players$college_name))])
```

```
[1] 12127
```

3.12.5 Number of Non-Missing Elements

```r
length(nfl_players$college_name[which(!is.na(nfl_players$college_name))])
```

```
[1] 7912
```

```
length(na.omit(nfl_players$college_name))
```

```
[1] 7912
```

3.13 Create New Variables

To create a new variable, use the following syntax:

```
players$newVar <- NA
```

Here is an example of creating a new variable:

```
players$newVar <- 1:nrow(players)
```

3.14 Recode Variables

Here is an example of recoding a variable:

```
players$oldVar1 <- NA
players$oldVar1[which(players$position == "QB")] <- "quarterback"
players$oldVar1[which(players$position == "RB")] <- "running back"
players$oldVar1[which(players$position == "WR")] <- "wide receiver"
players$oldVar1[which(players$position == "TE")] <- "tight end"

players$oldVar2 <- NA
players$oldVar2[which(players$age < 30)] <- "young"
players$oldVar2[which(players$age >= 30)] <- "old"
```

Recode multiple variables:

```
players %>%
  mutate(across(c(
    oldVar1:oldVar2),
    ~ case_match(
      .,
      c("quarterback","old","running back") ~ 0,
      c("wide receiver","tight end","young") ~ 1)))
```

```
   ID             name position age oldVar1 oldVar2
1   1      Ken Cussion       QB  40       0       0
2   2       Ben Sacked       QB  30       0       0
3   3  Chuck Downfield       QB  24       0       1
4   4      Ron Ingback       RB  20       0       1
5   5      Rhonda Ball       RB  18       0       1
6   6        Hugo Long       WR  23       1       1
7   7 Lionel Scrimmage       WR  27       1       1
8   8       Drew Blood       WR  32       1       0
9   9     Chase Emdown       WR  26       1       1
10 10      Justin Time       TE  23       1       1
11 11     Spike D'Ball       TE  NA       1      NA
12 12       Isac Ulooz       LB  37      NA       0
```

3.15 Rename Variables

```
players <- players %>%
  rename(
    newVar1 = oldVar1,
    newVar2 = oldVar2)
```

Using a vector of variable names:

```
varNamesFrom <- c("oldVar1","oldVar2")
varNamesTo <- c("newVar1","newVar2")

players <- players %>%
  rename_with(~ varNamesTo, all_of(varNamesFrom))
```

3.16 Convert the Types of Variables

One variable:

```
players$factorVar <- factor(players$ID)
players$numericVar <- as.numeric(players$age)
```

```
players$integerVar <- as.integer(players$newVar1)
players$characterVar <- as.character(players$newVar2)
```

Multiple variables:

```
players %>%
  mutate(across(c(
    ID,
    age),
    as.numeric))
```

```
   ID          name position age      newVar1 newVar2 factorVar numericVar
1   1     Ken Cussion      QB  40   quarterback     old        1        40
2   2      Ben Sacked      QB  30   quarterback     old        2        30
3   3  Chuck Downfield     QB  24   quarterback   young        3        24
4   4     Ron Ingback      RB  20  running back   young        4        20
5   5     Rhonda Ball      RB  18  running back   young        5        18
6   6       Hugo Long     WR  23 wide receiver   young        6        23
7   7 Lionel Scrimmage     WR  27 wide receiver   young        7        27
8   8      Drew Blood     WR  32 wide receiver     old        8        32
9   9    Chase Emdown     WR  26 wide receiver   young        9        26
10 10     Justin Time      TE  23     tight end   young       10        23
11 11    Spike D'Ball      TE  NA     tight end    <NA>       11        NA
12 12      Isac Ulooz      LB  37         <NA>     old       12        37
   integerVar characterVar
1          NA          old
2          NA          old
3          NA        young
4          NA        young
5          NA        young
6          NA        young
7          NA        young
8          NA          old
9          NA        young
10         NA        young
11         NA         <NA>
12         NA          old
```

```
players %>%
  mutate(across(
    age:newVar1,
    as.character))
```

```
   ID          name position  age      newVar1 newVar2 factorVar numericVar
```

```
1   1     Ken Cussion      QB   40   quarterback    old        1        40
2   2      Ben Sacked      QB   30   quarterback    old        2        30
3   3  Chuck Downfield     QB   24   quarterback   young       3        24
4   4     Ron Ingback      RB   20  running back   young       4        20
5   5     Rhonda Ball      RB   18  running back   young       5        18
6   6      Hugo Long      WR   23 wide receiver   young       6        23
7   7 Lionel Scrimmage     WR   27 wide receiver   young       7        27
8   8      Drew Blood      WR   32 wide receiver    old        8        32
9   9    Chase Emdown      WR   26 wide receiver   young       9        26
10 10     Justin Time      TE   23     tight end   young      10        23
11 11    Spike D'Ball      TE <NA>     tight end    <NA>      11        NA
12 12      Isac Ulooz      LB   37         <NA>    old       12        37
   integerVar characterVar
1          NA          old
2          NA          old
3          NA        young
4          NA        young
5          NA        young
6          NA        young
7          NA        young
8          NA          old
9          NA        young
10         NA        young
11         NA         <NA>
12         NA          old
```

```r
players %>%
  mutate(across(where(is.factor), as.character))
```

```
  ID          name position age      newVar1 newVar2 factorVar numericVar
1   1     Ken Cussion      QB  40   quarterback    old        1        40
2   2      Ben Sacked      QB  30   quarterback    old        2        30
3   3  Chuck Downfield     QB  24   quarterback   young       3        24
4   4     Ron Ingback      RB  20  running back   young       4        20
5   5     Rhonda Ball      RB  18  running back   young       5        18
6   6      Hugo Long      WR  23 wide receiver   young       6        23
7   7 Lionel Scrimmage     WR  27 wide receiver   young       7        27
8   8      Drew Blood      WR  32 wide receiver    old        8        32
9   9    Chase Emdown      WR  26 wide receiver   young       9        26
10 10     Justin Time      TE  23     tight end   young      10        23
11 11    Spike D'Ball      TE  NA     tight end    <NA>      11        NA
12 12      Isac Ulooz      LB  37         <NA>    old       12        37
   integerVar characterVar
1          NA          old
2          NA          old
```

```
3          NA        young
4          NA        young
5          NA        young
6          NA        young
7          NA        young
8          NA          old
9          NA        young
10         NA        young
11         NA         <NA>
12         NA          old
```

3.17 Merging/Joins

3.17.1 Overview

Merging (also called joining) merges two data objects using a shared set of variables called "keys." The keys are the variable(s) that uniquely identify each row (i.e., they account for the levels of nesting). In some data objects, the key might be the player's identification number (e.g., `player_id`). However, some data objects have multiple keys. For instance, in long form data objects, each participant may have multiple rows corresponding to multiple seasons. In this case, the keys may be `player_id` and `season`. If a participant has multiple rows corresponding to seasons and games/weeks, the keys are `player_id`, `season`, and `week`. In general, each row should have a value on each of the keys; there should be no missingness in the keys.

To merge two objects, the key(s) that will be used to match the records must be present in both objects. The keys are used to merge the variables in object 1 (`x`) with the variables in object 2 (`y`). Different merge types select different rows to merge.

Note: if the two objects include variables with the same name (apart from the keys), `R` will not know how you want each to appear in the merged object. So, it will add a suffix (e.g., `.x`, `.y`) to each common variable to indicate which object (i.e., object `x` or object `y`) the variable came from, where object `x` is the first object—i.e., the object to which object `y` (the second object) is merged. In general, apart from the keys, you should not include variables with the same name in two objects to be merged. To prevent this, either remove or rename the shared variable in one of the objects, or include the shared variable as a key. However, as described above, you should include it as a key ***only*** if it uniquely identifies each row in terms of levels of nesting.

3.17.2 Data Before Merging

Here are the data in the `players` object:

```
players
```

```
  ID          name position age     newVar1 newVar2 factorVar numericVar
1  1     Ken Cussion      QB  40   quarterback     old        1        40
2  2      Ben Sacked      QB  30   quarterback     old        2        30
3  3  Chuck Downfield      QB  24   quarterback   young        3        24
4  4     Ron Ingback      RB  20  running back   young        4        20
5  5     Rhonda Ball      RB  18  running back   young        5        18
6  6       Hugo Long      WR  23 wide receiver   young        6        23
7  7 Lionel Scrimmage      WR  27 wide receiver   young        7        27
8  8      Drew Blood      WR  32 wide receiver     old        8        32
9  9    Chase Emdown      WR  26 wide receiver   young        9        26
10 10     Justin Time      TE  23     tight end   young       10        23
11 11    Spike D'Ball      TE  NA     tight end    <NA>       11        NA
12 12      Isac Ulooz      LB  37         <NA>     old       12        37
   integerVar characterVar
1          NA          old
2          NA          old
3          NA        young
4          NA        young
5          NA        young
6          NA        young
7          NA        young
8          NA          old
9          NA        young
10         NA        young
11         NA         <NA>
12         NA          old
```

```
dim(players)
```

```
[1] 12 10
```

The data are structured in ID form. That is, every row in the dataset is uniquely identified by the variable, `ID`.

Here are the data in the `fantasyPoints` object:

```
fantasyPoints
```

```
  ID fantasyPoints
1  2           250
2  7           170
3 13            65
4 14            15
```

```r
dim(fantasyPoints)
```

```
[1] 4 2
```

3.17.3 Types of Joins

3.17.3.1 Visual Overview of Join Types

Below is a visual that depicts various types of merges/joins. Object `x` is the circle labeled as `x`. Object `y` is the circle labeled as `y`. The area of overlap in the Venn diagram indicates the rows on the keys that are shared between the two objects (e.g., the same `player_id`, `season`, and `week`). The non-overlapping area indicates the rows on the keys that are unique to each object. The shaded blue area indicates which rows (on the keys) are kept in the merged object from each of the two objects, when using each of the merge types. For instance, a left outer join keeps the shared rows and the rows that are unique to object `x`, but it drops the rows that are unique to object `y`.

Figure 3.1 Types of merges/joins

3.17.3.2 Full Outer Join

A full outer join includes all rows in `x` **or** `y`. It returns columns from `x` and `y`. Here is how to merge two data frames using a full outer join (i.e., "full join"):

```
fullJoinData <- full_join(
  players,
  fantasyPoints,
  by = "ID")

fullJoinData
```

```
   ID          name position age      newVar1 newVar2 factorVar numericVar
1   1     Ken Cussion      QB  40   quarterback     old        1        40
2   2      Ben Sacked      QB  30   quarterback     old        2        30
3   3  Chuck Downfield     QB  24   quarterback   young        3        24
4   4     Ron Ingback      RB  20  running back   young        4        20
5   5     Rhonda Ball      RB  18  running back   young        5        18
6   6       Hugo Long     WR  23 wide receiver   young        6        23
7   7 Lionel Scrimmage     WR  27 wide receiver   young        7        27
8   8      Drew Blood     WR  32 wide receiver     old        8        32
9   9    Chase Emdown     WR  26 wide receiver   young        9        26
10 10     Justin Time      TE  23    tight end   young       10        23
11 11    Spike D'Ball      TE  NA    tight end    <NA>       11        NA
12 12      Isac Ulooz      LB  37         <NA>     old       12        37
13 13           <NA>    <NA>  NA         <NA>    <NA>     <NA>        NA
14 14           <NA>    <NA>  NA         <NA>    <NA>     <NA>        NA
   integerVar characterVar fantasyPoints
1          NA          old            NA
2          NA          old           250
3          NA        young            NA
4          NA        young            NA
5          NA        young            NA
6          NA        young            NA
7          NA        young           170
8          NA          old            NA
9          NA        young            NA
10         NA        young            NA
11         NA         <NA>            NA
12         NA          old            NA
13         NA         <NA>            65
14         NA         <NA>            15
```

```r
dim(fullJoinData)
```

```
[1] 14 11
```

3.17.3.3 Left Outer Join

A left outer join includes all rows in `x`. It returns columns from `x` and `y`. Here is how to merge two data frames using a left outer join ("left join"):

```r
leftJoinData <- left_join(
  players,
  fantasyPoints,
  by = "ID")

leftJoinData
```

```
   ID          name position age      newVar1 newVar2 factorVar numericVar
1   1     Ken Cussion      QB  40   quarterback     old        1        40
2   2      Ben Sacked      QB  30   quarterback     old        2        30
3   3  Chuck Downfield     QB  24   quarterback   young        3        24
4   4     Ron Ingback      RB  20  running back   young        4        20
5   5     Rhonda Ball      RB  18  running back   young        5        18
6   6       Hugo Long     WR  23 wide receiver   young        6        23
7   7 Lionel Scrimmage     WR  27 wide receiver   young        7        27
8   8      Drew Blood     WR  32 wide receiver     old        8        32
9   9    Chase Emdown     WR  26 wide receiver   young        9        26
10 10     Justin Time      TE  23    tight end   young       10        23
11 11    Spike D'Ball      TE  NA    tight end    <NA>       11        NA
12 12      Isac Ulooz      LB  37         <NA>     old       12        37
   integerVar characterVar fantasyPoints
1          NA          old            NA
2          NA          old           250
3          NA        young            NA
4          NA        young            NA
5          NA        young            NA
6          NA        young            NA
7          NA        young           170
8          NA          old            NA
9          NA        young            NA
10         NA        young            NA
11         NA         <NA>            NA
12         NA          old            NA
```

```
dim(leftJoinData)
```

```
[1] 12 11
```

3.17.3.4 Right Outer Join

A right outer join includes all rows in `y`. It returns columns from `x` and `y`. Here is how to merge two data frames using a right outer join ("right join"):

```
rightJoinData <- right_join(
  players,
  fantasyPoints,
  by = "ID")

rightJoinData
```

```
  ID          name position age      newVar1 newVar2 factorVar numericVar
1  2      Ben Sacked      QB  30   quarterback     old        2        30
2  7 Lionel Scrimmage      WR  27 wide receiver   young        7        27
3 13            <NA>     <NA>  NA         <NA>    <NA>      <NA>        NA
4 14            <NA>     <NA>  NA         <NA>    <NA>      <NA>        NA
  integerVar characterVar fantasyPoints
1         NA          old           250
2         NA        young           170
3         NA         <NA>            65
4         NA         <NA>            15
```

```
dim(rightJoinData)
```

```
[1]  4 11
```

3.17.3.5 Inner Join

An inner join includes all rows that are in **both** `x` **and** `y`. An inner join will return one row of `x` for each matching row of `y`, and can duplicate values of records on either side (left or right) if `x` and `y` have more than one matching record. It returns columns from `x` and `y`. Here is how to merge two data frames using an inner join:

```
innerJoinData <- inner_join(
  players,
  fantasyPoints,
```

```
  by = "ID")

innerJoinData
```

```
  ID          name position age      newVar1 newVar2 factorVar numericVar
1  2      Ben Sacked      QB  30   quarterback     old        2        30
2  7 Lionel Scrimmage      WR  27 wide receiver   young        7        27
  integerVar characterVar fantasyPoints
1         NA          old           250
2         NA        young           170
```

```
dim(innerJoinData)
```

```
[1]  2 11
```

3.17.3.6 Semi Join

A semi join is a filter. A left semi join returns all rows from `x` **with** a match in `y`. That is, it filters out records from `x` that are not in `y`. Unlike an inner join, a left semi join will never duplicate rows of `x`, and it includes columns from only `x` (not from `y`). Here is how to merge two data frames using a left semi join:

```
semiJoinData <- semi_join(
  players,
  fantasyPoints,
  by = "ID")

semiJoinData
```

```
  ID          name position age      newVar1 newVar2 factorVar numericVar
1  2      Ben Sacked      QB  30   quarterback     old        2        30
2  7 Lionel Scrimmage      WR  27 wide receiver   young        7        27
  integerVar characterVar
1         NA          old
2         NA        young
```

```
dim(semiJoinData)
```

```
[1]  2 10
```

3.17.3.7 Anti Join

An anti join is a filter. A left anti join returns all rows from `x` **without** a match in `y`. That is, it filters out records from `x` that are in `y`. It returns columns from only `x` (not from `y`). Here is how to merge two data frames using a left anti join:

```
antiJoinData <- anti_join(
  players,
  fantasyPoints,
  by = "ID")

antiJoinData
```

```
   ID         name position age      newVar1 newVar2 factorVar numericVar
1   1    Ken Cussion      QB  40   quarterback     old        1        40
2   3 Chuck Downfield      QB  24   quarterback   young        3        24
3   4    Ron Ingback      RB  20  running back   young        4        20
4   5    Rhonda Ball      RB  18  running back   young        5        18
5   6      Hugo Long      WR  23 wide receiver   young        6        23
6   8     Drew Blood      WR  32 wide receiver     old        8        32
7   9   Chase Emdown      WR  26 wide receiver   young        9        26
8  10    Justin Time      TE  23     tight end   young       10        23
9  11   Spike D'Ball      TE  NA     tight end    <NA>       11        NA
10 12     Isac Ulooz      LB  37          <NA>     old       12        37
   integerVar characterVar
1          NA          old
2          NA        young
3          NA        young
4          NA        young
5          NA        young
6          NA          old
7          NA        young
8          NA        young
9          NA         <NA>
10         NA          old
```

```
dim(antiJoinData)
```

```
[1] 10 10
```

3.17.3.8 Cross Join

A cross join combines each row in `x` with each row in `y`.

```r
crossJoinData <- cross_join(
  players,
  fantasyPoints)

crossJoinData
```

```
   ID.x             name position age       newVar1 newVar2 factorVar
1     1      Ken Cussion       QB  40   quarterback     old         1
2     1      Ken Cussion       QB  40   quarterback     old         1
3     1      Ken Cussion       QB  40   quarterback     old         1
4     1      Ken Cussion       QB  40   quarterback     old         1
5     2       Ben Sacked       QB  30   quarterback     old         2
6     2       Ben Sacked       QB  30   quarterback     old         2
7     2       Ben Sacked       QB  30   quarterback     old         2
8     2       Ben Sacked       QB  30   quarterback     old         2
9     3  Chuck Downfield       QB  24   quarterback   young         3
10    3  Chuck Downfield       QB  24   quarterback   young         3
11    3  Chuck Downfield       QB  24   quarterback   young         3
12    3  Chuck Downfield       QB  24   quarterback   young         3
13    4      Ron Ingback       RB  20  running back   young         4
14    4      Ron Ingback       RB  20  running back   young         4
15    4      Ron Ingback       RB  20  running back   young         4
16    4      Ron Ingback       RB  20  running back   young         4
17    5      Rhonda Ball       RB  18  running back   young         5
18    5      Rhonda Ball       RB  18  running back   young         5
19    5      Rhonda Ball       RB  18  running back   young         5
20    5      Rhonda Ball       RB  18  running back   young         5
21    6        Hugo Long       WR  23 wide receiver   young         6
22    6        Hugo Long       WR  23 wide receiver   young         6
23    6        Hugo Long       WR  23 wide receiver   young         6
24    6        Hugo Long       WR  23 wide receiver   young         6
25    7 Lionel Scrimmage       WR  27 wide receiver   young         7
26    7 Lionel Scrimmage       WR  27 wide receiver   young         7
27    7 Lionel Scrimmage       WR  27 wide receiver   young         7
28    7 Lionel Scrimmage       WR  27 wide receiver   young         7
29    8       Drew Blood       WR  32 wide receiver     old         8
30    8       Drew Blood       WR  32 wide receiver     old         8
31    8       Drew Blood       WR  32 wide receiver     old         8
32    8       Drew Blood       WR  32 wide receiver     old         8
33    9     Chase Emdown       WR  26 wide receiver   young         9
34    9     Chase Emdown       WR  26 wide receiver   young         9
35    9     Chase Emdown       WR  26 wide receiver   young         9
36    9     Chase Emdown       WR  26 wide receiver   young         9
37   10      Justin Time       TE  23     tight end   young        10
38   10      Justin Time       TE  23     tight end   young        10
```

```
39   10      Justin Time       TE  23     tight end   young        10
40   10      Justin Time       TE  23     tight end   young        10
41   11     Spike D'Ball       TE  NA     tight end    <NA>        11
42   11     Spike D'Ball       TE  NA     tight end    <NA>        11
43   11     Spike D'Ball       TE  NA     tight end    <NA>        11
44   11     Spike D'Ball       TE  NA     tight end    <NA>        11
45   12       Isac Ulooz       LB  37          <NA>     old        12
46   12       Isac Ulooz       LB  37          <NA>     old        12
47   12       Isac Ulooz       LB  37          <NA>     old        12
48   12       Isac Ulooz       LB  37          <NA>     old        12
   numericVar integerVar characterVar ID.y fantasyPoints
1          40         NA          old    2           250
2          40         NA          old    7           170
3          40         NA          old   13            65
4          40         NA          old   14            15
5          30         NA          old    2           250
6          30         NA          old    7           170
7          30         NA          old   13            65
8          30         NA          old   14            15
9          24         NA        young    2           250
10         24         NA        young    7           170
11         24         NA        young   13            65
12         24         NA        young   14            15
13         20         NA        young    2           250
14         20         NA        young    7           170
15         20         NA        young   13            65
16         20         NA        young   14            15
17         18         NA        young    2           250
18         18         NA        young    7           170
19         18         NA        young   13            65
20         18         NA        young   14            15
21         23         NA        young    2           250
22         23         NA        young    7           170
23         23         NA        young   13            65
24         23         NA        young   14            15
25         27         NA        young    2           250
26         27         NA        young    7           170
27         27         NA        young   13            65
28         27         NA        young   14            15
29         32         NA          old    2           250
30         32         NA          old    7           170
31         32         NA          old   13            65
32         32         NA          old   14            15
33         26         NA        young    2           250
34         26         NA        young    7           170
```

```
35          26          NA         young   13            65
36          26          NA         young   14            15
37          23          NA         young    2           250
38          23          NA         young    7           170
39          23          NA         young   13            65
40          23          NA         young   14            15
41          NA          NA          <NA>    2           250
42          NA          NA          <NA>    7           170
43          NA          NA          <NA>   13            65
44          NA          NA          <NA>   14            15
45          37          NA           old    2           250
46          37          NA           old    7           170
47          37          NA           old   13            65
48          37          NA           old   14            15
```

```
dim(crossJoinData)
```

```
[1] 48 12
```

3.18 Transform Data from Long to Wide

Depending on the analysis, it may be important to restructure the data to be in long or wide form. When the data are in wide form, each player has only one row. When the data are in long form, each player has multiple rows—e.g., a row for each game. The data structure is called wide or long form because a dataset in wide form has more columns and fewer rows (i.e., it appears wider and shorter), whereas a dataset in long form has more rows and fewer columns (i.e., it appears narrower and taller).

Here are the data in the `nfl_actualStats_offense_weekly` object. The data are structured in "player-season-week form". That is, every row in the dataset is uniquely identified by the variables, `player_id`, `season`, and `week`. This is an example of long form, because each player has multiple rows.

Original data:

```
dataLong <- nfl_actualStats_offense_weekly %>%
  select(player_id, player_display_name, season, week, fantasy_points)

dim(dataLong)
```

```
[1] 129739      5
```

```
names(dataLong)
```

```
[1] "player_id"           "player_display_name" "season"
[4] "week"                "fantasy_points"
```

Below, we widen the data by two variables (`season` and `week`), using `tidyverse`, so that the data are now in "player form" (where each row is uniquely identified by the `player_id` variable):

```
dataWide <- dataLong %>%
  pivot_wider(
    names_from = c(season, week),
    names_glue = "{.value}_{season}_week{week}",
    values_from = fantasy_points)

dim(dataWide)
```

```
[1] 4021  530
```

```
names(dataWide)
```

```
  [1] "player_id"                  "player_display_name"
  [3] "fantasy_points_1999_week1"  "fantasy_points_1999_week2"
  [5] "fantasy_points_1999_week4"  "fantasy_points_1999_week7"
  [7] "fantasy_points_1999_week8"  "fantasy_points_1999_week9"
  [9] "fantasy_points_1999_week10" "fantasy_points_1999_week11"
 [11] "fantasy_points_1999_week12" "fantasy_points_1999_week13"
 [13] "fantasy_points_1999_week14" "fantasy_points_1999_week15"
 [15] "fantasy_points_1999_week16" "fantasy_points_1999_week5"
 [17] "fantasy_points_1999_week6"  "fantasy_points_1999_week17"
 [19] "fantasy_points_1999_week18" "fantasy_points_1999_week3"
 [21] "fantasy_points_1999_week19" "fantasy_points_1999_week20"
 [23] "fantasy_points_1999_week21" "fantasy_points_2000_week1"
 [25] "fantasy_points_2000_week12" "fantasy_points_2000_week14"
 [27] "fantasy_points_2000_week15" "fantasy_points_2000_week6"
 [29] "fantasy_points_2000_week10" "fantasy_points_2000_week4"
 [31] "fantasy_points_2000_week5"  "fantasy_points_2000_week7"
 [33] "fantasy_points_2000_week8"  "fantasy_points_2000_week9"
 [35] "fantasy_points_2000_week11" "fantasy_points_2000_week13"
 [37] "fantasy_points_2000_week2"  "fantasy_points_2000_week16"
 [39] "fantasy_points_2000_week17" "fantasy_points_2000_week3"
 [41] "fantasy_points_2000_week18" "fantasy_points_2000_week19"
 [43] "fantasy_points_2000_week21" "fantasy_points_2000_week20"
```

```
 [45] "fantasy_points_2001_week15" "fantasy_points_2001_week17"
 [47] "fantasy_points_2001_week1"  "fantasy_points_2001_week3"
 [49] "fantasy_points_2001_week4"  "fantasy_points_2001_week5"
 [51] "fantasy_points_2001_week6"  "fantasy_points_2001_week9"
 [53] "fantasy_points_2001_week11" "fantasy_points_2001_week12"
 [55] "fantasy_points_2001_week13" "fantasy_points_2001_week14"
 [57] "fantasy_points_2001_week16" "fantasy_points_2001_week2"
 [59] "fantasy_points_2001_week7"  "fantasy_points_2001_week8"
 [61] "fantasy_points_2001_week10" "fantasy_points_2001_week19"
 [63] "fantasy_points_2001_week18" "fantasy_points_2001_week20"
 [65] "fantasy_points_2001_week21" "fantasy_points_2002_week3"
 [67] "fantasy_points_2002_week1"  "fantasy_points_2002_week2"
 [69] "fantasy_points_2002_week4"  "fantasy_points_2002_week6"
 [71] "fantasy_points_2002_week7"  "fantasy_points_2002_week8"
 [73] "fantasy_points_2002_week9"  "fantasy_points_2002_week5"
 [75] "fantasy_points_2002_week10" "fantasy_points_2002_week11"
 [77] "fantasy_points_2002_week12" "fantasy_points_2002_week13"
 [79] "fantasy_points_2002_week14" "fantasy_points_2002_week15"
 [81] "fantasy_points_2002_week16" "fantasy_points_2002_week17"
 [83] "fantasy_points_2002_week19" "fantasy_points_2002_week20"
 [85] "fantasy_points_2002_week21" "fantasy_points_2002_week18"
 [87] "fantasy_points_2003_week2"  "fantasy_points_2003_week4"
 [89] "fantasy_points_2003_week5"  "fantasy_points_2003_week7"
 [91] "fantasy_points_2003_week8"  "fantasy_points_2003_week9"
 [93] "fantasy_points_2003_week10" "fantasy_points_2003_week11"
 [95] "fantasy_points_2003_week13" "fantasy_points_2003_week14"
 [97] "fantasy_points_2003_week15" "fantasy_points_2003_week17"
 [99] "fantasy_points_2003_week1"  "fantasy_points_2003_week3"
[101] "fantasy_points_2003_week6"  "fantasy_points_2003_week12"
[103] "fantasy_points_2003_week16" "fantasy_points_2003_week18"
[105] "fantasy_points_2003_week19" "fantasy_points_2003_week20"
[107] "fantasy_points_2003_week21" "fantasy_points_2004_week2"
[109] "fantasy_points_2004_week5"  "fantasy_points_2004_week6"
[111] "fantasy_points_2004_week10" "fantasy_points_2004_week16"
[113] "fantasy_points_2004_week17" "fantasy_points_2004_week1"
[115] "fantasy_points_2004_week3"  "fantasy_points_2004_week7"
[117] "fantasy_points_2004_week8"  "fantasy_points_2004_week9"
[119] "fantasy_points_2004_week11" "fantasy_points_2004_week12"
[121] "fantasy_points_2004_week13" "fantasy_points_2004_week14"
[123] "fantasy_points_2004_week15" "fantasy_points_2004_week4"
[125] "fantasy_points_2004_week19" "fantasy_points_2004_week20"
[127] "fantasy_points_2004_week21" "fantasy_points_2004_week18"
[129] "fantasy_points_2005_week1"  "fantasy_points_2005_week2"
[131] "fantasy_points_2005_week3"  "fantasy_points_2005_week4"
[133] "fantasy_points_2005_week6"  "fantasy_points_2005_week7"
```

```
[135] "fantasy_points_2005_week8"  "fantasy_points_2005_week10"
[137] "fantasy_points_2005_week11" "fantasy_points_2005_week13"
[139] "fantasy_points_2005_week14" "fantasy_points_2005_week15"
[141] "fantasy_points_2005_week16" "fantasy_points_2005_week17"
[143] "fantasy_points_2005_week5"  "fantasy_points_2005_week9"
[145] "fantasy_points_2005_week12" "fantasy_points_2005_week18"
[147] "fantasy_points_2005_week19" "fantasy_points_2005_week20"
[149] "fantasy_points_2005_week21" "fantasy_points_2006_week4"
[151] "fantasy_points_2006_week1"  "fantasy_points_2006_week2"
[153] "fantasy_points_2006_week3"  "fantasy_points_2006_week10"
[155] "fantasy_points_2006_week11" "fantasy_points_2006_week12"
[157] "fantasy_points_2006_week13" "fantasy_points_2006_week14"
[159] "fantasy_points_2006_week5"  "fantasy_points_2006_week6"
[161] "fantasy_points_2006_week7"  "fantasy_points_2006_week15"
[163] "fantasy_points_2006_week16" "fantasy_points_2006_week17"
[165] "fantasy_points_2006_week8"  "fantasy_points_2006_week9"
[167] "fantasy_points_2006_week18" "fantasy_points_2006_week19"
[169] "fantasy_points_2006_week20" "fantasy_points_2006_week21"
[171] "fantasy_points_2007_week1"  "fantasy_points_2007_week2"
[173] "fantasy_points_2007_week3"  "fantasy_points_2007_week5"
[175] "fantasy_points_2007_week9"  "fantasy_points_2007_week16"
[177] "fantasy_points_2007_week17" "fantasy_points_2007_week14"
[179] "fantasy_points_2007_week4"  "fantasy_points_2007_week6"
[181] "fantasy_points_2007_week7"  "fantasy_points_2007_week8"
[183] "fantasy_points_2007_week10" "fantasy_points_2007_week11"
[185] "fantasy_points_2007_week12" "fantasy_points_2007_week13"
[187] "fantasy_points_2007_week15" "fantasy_points_2007_week19"
[189] "fantasy_points_2007_week21" "fantasy_points_2007_week18"
[191] "fantasy_points_2007_week20" "fantasy_points_2008_week8"
[193] "fantasy_points_2008_week1"  "fantasy_points_2008_week2"
[195] "fantasy_points_2008_week3"  "fantasy_points_2008_week4"
[197] "fantasy_points_2008_week5"  "fantasy_points_2008_week6"
[199] "fantasy_points_2008_week7"  "fantasy_points_2008_week14"
[201] "fantasy_points_2008_week10" "fantasy_points_2008_week11"
[203] "fantasy_points_2008_week12" "fantasy_points_2008_week13"
[205] "fantasy_points_2008_week15" "fantasy_points_2008_week16"
[207] "fantasy_points_2008_week17" "fantasy_points_2008_week9"
[209] "fantasy_points_2008_week19" "fantasy_points_2008_week18"
[211] "fantasy_points_2008_week20" "fantasy_points_2008_week21"
[213] "fantasy_points_2009_week9"  "fantasy_points_2009_week11"
[215] "fantasy_points_2009_week2"  "fantasy_points_2009_week3"
[217] "fantasy_points_2009_week5"  "fantasy_points_2009_week7"
[219] "fantasy_points_2009_week12" "fantasy_points_2009_week13"
[221] "fantasy_points_2009_week14" "fantasy_points_2009_week15"
[223] "fantasy_points_2009_week16" "fantasy_points_2009_week17"
```

```
[225] "fantasy_points_2009_week1"  "fantasy_points_2009_week4"
[227] "fantasy_points_2009_week8"  "fantasy_points_2009_week6"
[229] "fantasy_points_2009_week10" "fantasy_points_2009_week18"
[231] "fantasy_points_2009_week19" "fantasy_points_2009_week20"
[233] "fantasy_points_2009_week21" "fantasy_points_2010_week2"
[235] "fantasy_points_2010_week3"  "fantasy_points_2010_week4"
[237] "fantasy_points_2010_week1"  "fantasy_points_2010_week7"
[239] "fantasy_points_2010_week17" "fantasy_points_2010_week6"
[241] "fantasy_points_2010_week8"  "fantasy_points_2010_week10"
[243] "fantasy_points_2010_week13" "fantasy_points_2010_week14"
[245] "fantasy_points_2010_week15" "fantasy_points_2010_week16"
[247] "fantasy_points_2010_week5"  "fantasy_points_2010_week20"
[249] "fantasy_points_2010_week12" "fantasy_points_2010_week11"
[251] "fantasy_points_2010_week18" "fantasy_points_2010_week19"
[253] "fantasy_points_2010_week21" "fantasy_points_2010_week9"
[255] "fantasy_points_2011_week17" "fantasy_points_2011_week13"
[257] "fantasy_points_2011_week14" "fantasy_points_2011_week16"
[259] "fantasy_points_2011_week15" "fantasy_points_2011_week1"
[261] "fantasy_points_2011_week2"  "fantasy_points_2011_week3"
[263] "fantasy_points_2011_week4"  "fantasy_points_2011_week5"
[265] "fantasy_points_2011_week6"  "fantasy_points_2011_week7"
[267] "fantasy_points_2011_week9"  "fantasy_points_2011_week10"
[269] "fantasy_points_2011_week11" "fantasy_points_2011_week12"
[271] "fantasy_points_2011_week19" "fantasy_points_2011_week8"
[273] "fantasy_points_2011_week18" "fantasy_points_2011_week20"
[275] "fantasy_points_2011_week21" "fantasy_points_2012_week12"
[277] "fantasy_points_2012_week13" "fantasy_points_2012_week2"
[279] "fantasy_points_2012_week3"  "fantasy_points_2012_week4"
[281] "fantasy_points_2012_week5"  "fantasy_points_2012_week7"
[283] "fantasy_points_2012_week8"  "fantasy_points_2012_week9"
[285] "fantasy_points_2012_week11" "fantasy_points_2012_week16"
[287] "fantasy_points_2012_week1"  "fantasy_points_2012_week6"
[289] "fantasy_points_2012_week10" "fantasy_points_2012_week14"
[291] "fantasy_points_2012_week15" "fantasy_points_2012_week17"
[293] "fantasy_points_2012_week19" "fantasy_points_2012_week20"
[295] "fantasy_points_2012_week21" "fantasy_points_2012_week18"
[297] "fantasy_points_2013_week1"  "fantasy_points_2013_week2"
[299] "fantasy_points_2013_week3"  "fantasy_points_2013_week4"
[301] "fantasy_points_2013_week5"  "fantasy_points_2013_week7"
[303] "fantasy_points_2013_week8"  "fantasy_points_2013_week9"
[305] "fantasy_points_2013_week10" "fantasy_points_2013_week11"
[307] "fantasy_points_2013_week12" "fantasy_points_2013_week13"
[309] "fantasy_points_2013_week14" "fantasy_points_2013_week15"
[311] "fantasy_points_2013_week16" "fantasy_points_2013_week17"
[313] "fantasy_points_2013_week6"  "fantasy_points_2013_week19"
```

```
[315] "fantasy_points_2013_week20" "fantasy_points_2013_week21"
[317] "fantasy_points_2013_week18" "fantasy_points_2014_week3"
[319] "fantasy_points_2014_week4"  "fantasy_points_2014_week16"
[321] "fantasy_points_2014_week17" "fantasy_points_2014_week1"
[323] "fantasy_points_2014_week2"  "fantasy_points_2014_week5"
[325] "fantasy_points_2014_week6"  "fantasy_points_2014_week7"
[327] "fantasy_points_2014_week8"  "fantasy_points_2014_week9"
[329] "fantasy_points_2014_week10" "fantasy_points_2014_week11"
[331] "fantasy_points_2014_week12" "fantasy_points_2014_week13"
[333] "fantasy_points_2014_week14" "fantasy_points_2014_week15"
[335] "fantasy_points_2014_week19" "fantasy_points_2014_week20"
[337] "fantasy_points_2014_week21" "fantasy_points_2014_week18"
[339] "fantasy_points_2015_week4"  "fantasy_points_2015_week5"
[341] "fantasy_points_2015_week11" "fantasy_points_2015_week12"
[343] "fantasy_points_2015_week13" "fantasy_points_2015_week14"
[345] "fantasy_points_2015_week15" "fantasy_points_2015_week16"
[347] "fantasy_points_2015_week1"  "fantasy_points_2015_week2"
[349] "fantasy_points_2015_week3"  "fantasy_points_2015_week6"
[351] "fantasy_points_2015_week8"  "fantasy_points_2015_week9"
[353] "fantasy_points_2015_week10" "fantasy_points_2015_week17"
[355] "fantasy_points_2015_week19" "fantasy_points_2015_week20"
[357] "fantasy_points_2015_week21" "fantasy_points_2015_week7"
[359] "fantasy_points_2015_week18" "fantasy_points_2016_week5"
[361] "fantasy_points_2016_week6"  "fantasy_points_2016_week7"
[363] "fantasy_points_2016_week8"  "fantasy_points_2016_week10"
[365] "fantasy_points_2016_week11" "fantasy_points_2016_week12"
[367] "fantasy_points_2016_week13" "fantasy_points_2016_week14"
[369] "fantasy_points_2016_week15" "fantasy_points_2016_week16"
[371] "fantasy_points_2016_week17" "fantasy_points_2016_week19"
[373] "fantasy_points_2016_week20" "fantasy_points_2016_week21"
[375] "fantasy_points_2016_week1"  "fantasy_points_2016_week2"
[377] "fantasy_points_2016_week3"  "fantasy_points_2016_week4"
[379] "fantasy_points_2016_week9"  "fantasy_points_2016_week18"
[381] "fantasy_points_2017_week1"  "fantasy_points_2017_week2"
[383] "fantasy_points_2017_week3"  "fantasy_points_2017_week4"
[385] "fantasy_points_2017_week5"  "fantasy_points_2017_week6"
[387] "fantasy_points_2017_week7"  "fantasy_points_2017_week8"
[389] "fantasy_points_2017_week10" "fantasy_points_2017_week11"
[391] "fantasy_points_2017_week12" "fantasy_points_2017_week13"
[393] "fantasy_points_2017_week14" "fantasy_points_2017_week15"
[395] "fantasy_points_2017_week16" "fantasy_points_2017_week17"
[397] "fantasy_points_2017_week19" "fantasy_points_2017_week20"
[399] "fantasy_points_2017_week21" "fantasy_points_2017_week9"
[401] "fantasy_points_2017_week18" "fantasy_points_2018_week1"
[403] "fantasy_points_2018_week2"  "fantasy_points_2018_week3"
```

```
[405] "fantasy_points_2018_week4"  "fantasy_points_2018_week5"
[407] "fantasy_points_2018_week6"  "fantasy_points_2018_week7"
[409] "fantasy_points_2018_week8"  "fantasy_points_2018_week9"
[411] "fantasy_points_2018_week10" "fantasy_points_2018_week12"
[413] "fantasy_points_2018_week13" "fantasy_points_2018_week14"
[415] "fantasy_points_2018_week15" "fantasy_points_2018_week16"
[417] "fantasy_points_2018_week17" "fantasy_points_2018_week19"
[419] "fantasy_points_2018_week20" "fantasy_points_2018_week21"
[421] "fantasy_points_2018_week11" "fantasy_points_2018_week18"
[423] "fantasy_points_2019_week1"  "fantasy_points_2019_week2"
[425] "fantasy_points_2019_week3"  "fantasy_points_2019_week4"
[427] "fantasy_points_2019_week5"  "fantasy_points_2019_week6"
[429] "fantasy_points_2019_week7"  "fantasy_points_2019_week8"
[431] "fantasy_points_2019_week9"  "fantasy_points_2019_week11"
[433] "fantasy_points_2019_week12" "fantasy_points_2019_week13"
[435] "fantasy_points_2019_week14" "fantasy_points_2019_week15"
[437] "fantasy_points_2019_week16" "fantasy_points_2019_week17"
[439] "fantasy_points_2019_week18" "fantasy_points_2019_week10"
[441] "fantasy_points_2019_week19" "fantasy_points_2019_week20"
[443] "fantasy_points_2019_week21" "fantasy_points_2020_week1"
[445] "fantasy_points_2020_week2"  "fantasy_points_2020_week3"
[447] "fantasy_points_2020_week4"  "fantasy_points_2020_week5"
[449] "fantasy_points_2020_week6"  "fantasy_points_2020_week7"
[451] "fantasy_points_2020_week8"  "fantasy_points_2020_week9"
[453] "fantasy_points_2020_week10" "fantasy_points_2020_week11"
[455] "fantasy_points_2020_week12" "fantasy_points_2020_week14"
[457] "fantasy_points_2020_week15" "fantasy_points_2020_week16"
[459] "fantasy_points_2020_week17" "fantasy_points_2020_week18"
[461] "fantasy_points_2020_week19" "fantasy_points_2020_week20"
[463] "fantasy_points_2020_week21" "fantasy_points_2020_week13"
[465] "fantasy_points_2021_week1"  "fantasy_points_2021_week2"
[467] "fantasy_points_2021_week3"  "fantasy_points_2021_week4"
[469] "fantasy_points_2021_week5"  "fantasy_points_2021_week6"
[471] "fantasy_points_2021_week7"  "fantasy_points_2021_week8"
[473] "fantasy_points_2021_week10" "fantasy_points_2021_week11"
[475] "fantasy_points_2021_week12" "fantasy_points_2021_week13"
[477] "fantasy_points_2021_week14" "fantasy_points_2021_week15"
[479] "fantasy_points_2021_week16" "fantasy_points_2021_week17"
[481] "fantasy_points_2021_week18" "fantasy_points_2021_week19"
[483] "fantasy_points_2021_week20" "fantasy_points_2021_week9"
[485] "fantasy_points_2021_week21" "fantasy_points_2021_week22"
[487] "fantasy_points_2022_week1"  "fantasy_points_2022_week2"
[489] "fantasy_points_2022_week3"  "fantasy_points_2022_week4"
[491] "fantasy_points_2022_week5"  "fantasy_points_2022_week6"
[493] "fantasy_points_2022_week7"  "fantasy_points_2022_week8"
```

```
[495] "fantasy_points_2022_week9"  "fantasy_points_2022_week10"
[497] "fantasy_points_2022_week12" "fantasy_points_2022_week13"
[499] "fantasy_points_2022_week14" "fantasy_points_2022_week15"
[501] "fantasy_points_2022_week16" "fantasy_points_2022_week17"
[503] "fantasy_points_2022_week18" "fantasy_points_2022_week19"
[505] "fantasy_points_2022_week11" "fantasy_points_2022_week20"
[507] "fantasy_points_2022_week21" "fantasy_points_2022_week22"
[509] "fantasy_points_2023_week1"  "fantasy_points_2023_week4"
[511] "fantasy_points_2023_week7"  "fantasy_points_2023_week11"
[513] "fantasy_points_2023_week14" "fantasy_points_2023_week16"
[515] "fantasy_points_2023_week13" "fantasy_points_2023_week15"
[517] "fantasy_points_2023_week17" "fantasy_points_2023_week19"
[519] "fantasy_points_2023_week2"  "fantasy_points_2023_week3"
[521] "fantasy_points_2023_week5"  "fantasy_points_2023_week6"
[523] "fantasy_points_2023_week8"  "fantasy_points_2023_week12"
[525] "fantasy_points_2023_week18" "fantasy_points_2023_week10"
[527] "fantasy_points_2023_week21" "fantasy_points_2023_week22"
[529] "fantasy_points_2023_week9"  "fantasy_points_2023_week20"
```

3.19 Transform Data from Wide to Long

Conversely, we can also restructure data from wide to long.

Original data:

```r
dataWide <- nfl_actualStats_offense_weekly %>%
  select(player_id, player_display_name, season, week, recent_team, opponent_team)

dim(dataWide)
```

```
[1] 129739      6
```

```r
names(dataWide)
```

```
[1] "player_id"           "player_display_name" "season"
[4] "week"                "recent_team"         "opponent_team"
```

Data in long form, transformed from wide form using `tidyverse`:

```
dataLong <- dataWide %>%
  pivot_longer(
    cols = c(recent_team, opponent_team),
    names_to = "role",
    values_to = "team")

dim(dataLong)
```

```
[1] 259478      6
```

```
names(dataLong)
```

```
[1] "player_id"           "player_display_name" "season"
[4] "week"                "role"                "team"
```

3.20 Loops

If you want to perform the same computation multiple times, it can be faster to do it in a loop compared to writing out the same computation many times. For instance, here is a loop that runs from 1 to 12 (the number of players in the `players` object), incrementing by 1 after each iteration. The loop prints each element of a vector (i.e., the player's name) and the loop index (`i`) that indicates where the loop is in terms of its iterations:

```
for(i in 1:length(players$ID)){
  print(paste("The loop is at index:", i, sep = " "))
  print(paste("My favorite player is:", players$name[i], sep = " "))
}
```

```
[1] "The loop is at index: 1"
[1] "My favorite player is: Ken Cussion"
[1] "The loop is at index: 2"
[1] "My favorite player is: Ben Sacked"
[1] "The loop is at index: 3"
[1] "My favorite player is: Chuck Downfield"
[1] "The loop is at index: 4"
[1] "My favorite player is: Ron Ingback"
[1] "The loop is at index: 5"
[1] "My favorite player is: Rhonda Ball"
```

```
[1] "The loop is at index: 6"
[1] "My favorite player is: Hugo Long"
[1] "The loop is at index: 7"
[1] "My favorite player is: Lionel Scrimmage"
[1] "The loop is at index: 8"
[1] "My favorite player is: Drew Blood"
[1] "The loop is at index: 9"
[1] "My favorite player is: Chase Emdown"
[1] "The loop is at index: 10"
[1] "My favorite player is: Justin Time"
[1] "The loop is at index: 11"
[1] "My favorite player is: Spike D'Ball"
[1] "The loop is at index: 12"
[1] "My favorite player is: Isac Ulooz"
```

3.21 Calculations

3.21.1 Historical Actual Player Statistics

In addition to week-by-week actual player statistics, we can also compute historical actual player statistics as a function of different timeframes, including season-by-season and career statistics.

3.21.1.1 Career Statistics

First, we can compute the players' career statistics using the `calculate_player_stats()`, `calculate_player_stats_def()`, and `calculate_player_stats_kicking()` functions from the `nflfastR` package for offensive players, defensive players, and kickers, respectively.

> **Note 3: Calculating players' career statistics**
>
> Note: the following code takes a while to run.

```r
nfl_actualStats_offense_career <- nflfastR::calculate_player_stats(
  nfl_pbp,
  weekly = FALSE)

nfl_actualStats_defense_career <- nflfastR::calculate_player_stats_def(
  nfl_pbp,
```

```
  weekly = FALSE)

nfl_actualStats_kicking_career <- nflfastR::calculate_player_stats_kicking(
  nfl_pbp,
  weekly = FALSE)
```

3.21.1.2 Season-by-Season Statistics

Second, we can compute the players' season-by-season statistics.

```
seasons <- unique(nfl_pbp$season)

nfl_pbp_seasonalList <- list()
nfl_actualStats_offense_seasonalList <- list()
nfl_actualStats_defense_seasonalList <- list()
nfl_actualStats_kicking_seasonalList <- list()
```

> **Note 4: Calculating players' season-by-season statistics**
>
> Note: the following code takes a while to run.

```
pb <- txtProgressBar(
  min = 0,
  max = length(seasons),
  style = 3)

for(i in 1:length(seasons)){
  # Subset play-by-play data by season
  nfl_pbp_seasonalList[[i]] <- nfl_pbp %>%
    filter(season == seasons[i])

  # Compute actual statistics by season
  nfl_actualStats_offense_seasonalList[[i]] <-
    nflfastR::calculate_player_stats(
      nfl_pbp_seasonalList[[i]],
      weekly = FALSE)

  nfl_actualStats_defense_seasonalList[[i]] <-
    nflfastR::calculate_player_stats_def(
      nfl_pbp_seasonalList[[i]],
      weekly = FALSE)

  nfl_actualStats_kicking_seasonalList[[i]] <-
```

```
    nflfastR::calculate_player_stats_kicking(
      nfl_pbp_seasonalList[[i]],
      weekly = FALSE)

  nfl_actualStats_offense_seasonalList[[i]]$season <- seasons[i]
  nfl_actualStats_defense_seasonalList[[i]]$season <- seasons[i]
  nfl_actualStats_kicking_seasonalList[[i]]$season <- seasons[i]

  print(
    paste("Completed computing projections for season: ", seasons[i], sep = ""))

  # Update the progress bar
  setTxtProgressBar(pb, i)
}

# Close the progress bar
close(pb)

nfl_actualStats_offense_seasonal <- nfl_actualStats_offense_seasonalList %>%
  bind_rows()
nfl_actualStats_defense_seasonal <- nfl_actualStats_defense_seasonalList %>%
  bind_rows()
nfl_actualStats_kicking_seasonal <- nfl_actualStats_kicking_seasonalList %>%
  bind_rows()
```

3.21.1.3 Week-by-Week Statistics

We already load players' week-by-week statistics above. Nevertheless, we could compute players' weekly statistics from the play-by-play data using the following syntax:

```
nfl_actualStats_offense_weekly <- nflfastR::calculate_player_stats(
  nfl_pbp,
  weekly = TRUE)

nfl_actualStats_defense_weekly <- nflfastR::calculate_player_stats_def(
  nfl_pbp,
  weekly = TRUE)

nfl_actualStats_kicking_weekly <- nflfastR::calculate_player_stats_kicking(
  nfl_pbp,
  weekly = TRUE)
```

3.21.2 Historical Actual Fantasy Points

Specify scoring settings:

3.21.2.1 Weekly

3.21.2.2 Seasonal

3.21.2.3 Career

3.21.3 Player Age

```
# Reshape from wide to long format
nfl_actualStats_offense_weekly_long <- nfl_actualStats_offense_weekly %>%
  pivot_longer(
    cols = c(recent_team, opponent_team),
    names_to = "role",
    values_to = "team")

# Perform separate inner join operations for the home_team and away_team
nfl_actualStats_offense_weekly_home <- inner_join(
  nfl_actualStats_offense_weekly_long,
  nfl_schedules,
  by = c("season","week","team" = "home_team")) %>%
  mutate(home_away = "home_team")

nfl_actualStats_offense_weekly_away <- inner_join(
  nfl_actualStats_offense_weekly_long,
  nfl_schedules,
  by = c("season","week","team" = "away_team")) %>%
  mutate(home_away = "away_team")

nfl_actualStats_defense_weekly_home <- inner_join(
  nfl_actualStats_defense_weekly,
  nfl_schedules,
  by = c("season","week","team" = "home_team")) %>%
  mutate(home_away = "home_team")

nfl_actualStats_defense_weekly_away <- inner_join(
  nfl_actualStats_defense_weekly,
  nfl_schedules,
  by = c("season","week","team" = "away_team")) %>%
  mutate(home_away = "away_team")
```

```
nfl_actualStats_kicking_weekly_home <- inner_join(
  nfl_actualStats_kicking_weekly,
  nfl_schedules,
  by = c("season","week","team" = "home_team")) %>%
  mutate(home_away = "home_team")

nfl_actualStats_kicking_weekly_away <- inner_join(
  nfl_actualStats_kicking_weekly,
  nfl_schedules,
  by = c("season","week","team" = "away_team")) %>%
  mutate(home_away = "away_team")

# Combine the results of the join operations
nfl_actualStats_offense_weekly_schedules_long <- bind_rows(
  nfl_actualStats_offense_weekly_home,
  nfl_actualStats_offense_weekly_away)

nfl_actualStats_defense_weekly_schedules_long <- bind_rows(
  nfl_actualStats_defense_weekly_home,
  nfl_actualStats_defense_weekly_away)

nfl_actualStats_kicking_weekly_schedules_long <- bind_rows(
  nfl_actualStats_kicking_weekly_home,
  nfl_actualStats_kicking_weekly_away)

# Reshape from long to wide
player_game_gameday_offense <- nfl_actualStats_offense_weekly_schedules_long %>%
  distinct(player_id, season, week, game_id, home_away, team, gameday) %>% #, .keep_all = TRUE
  pivot_wider(
    names_from = home_away,
    values_from = team)

player_game_gameday_defense <- nfl_actualStats_defense_weekly_schedules_long %>%
  distinct(player_id, season, week, game_id, home_away, team, gameday) %>% #, .keep_all = TRUE
  pivot_wider(
    names_from = home_away,
    values_from = team)

player_game_gameday_kicking <- nfl_actualStats_kicking_weekly_schedules_long %>%
  distinct(player_id, season, week, game_id, home_away, team, gameday) %>% #, .keep_all = TRUE
  pivot_wider(
    names_from = home_away,
    values_from = team)
```

```
# Merge player birthdate and the game date
player_game_birthdate_gameday_offense <- left_join(
  player_game_gameday_offense,
  unique(nfl_players[,c("gsis_id","birth_date")]),
  by = c("player_id" = "gsis_id")
)

player_game_birthdate_gameday_defense <- left_join(
  player_game_gameday_defense,
  unique(nfl_players[,c("gsis_id","birth_date")]),
  by = c("player_id" = "gsis_id")
)

player_game_birthdate_gameday_kicking <- left_join(
  player_game_gameday_kicking,
  unique(nfl_players[,c("gsis_id","birth_date")]),
  by = c("player_id" = "gsis_id")
)

player_game_birthdate_gameday_offense$birth_date <- ymd(player_game_birthdate_gameday_offense$birt
player_game_birthdate_gameday_offense$gameday <- ymd(player_game_birthdate_gameday_offense$gameday

player_game_birthdate_gameday_defense$birth_date <- ymd(player_game_birthdate_gameday_defense$birt
player_game_birthdate_gameday_defense$gameday <- ymd(player_game_birthdate_gameday_defense$gameday

player_game_birthdate_gameday_kicking$birth_date <- ymd(player_game_birthdate_gameday_kicking$birt
player_game_birthdate_gameday_kicking$gameday <- ymd(player_game_birthdate_gameday_kicking$gameday

# Calculate player's age for a given week as the difference between their birthdate and the game d
player_game_birthdate_gameday_offense$age <- interval(
  start = player_game_birthdate_gameday_offense$birth_date,
  end = player_game_birthdate_gameday_offense$gameday
) %>%
  time_length(unit = "years")

player_game_birthdate_gameday_defense$age <- interval(
  start = player_game_birthdate_gameday_defense$birth_date,
  end = player_game_birthdate_gameday_defense$gameday
) %>%
  time_length(unit = "years")

player_game_birthdate_gameday_kicking$age <- interval(
  start = player_game_birthdate_gameday_kicking$birth_date,
  end = player_game_birthdate_gameday_kicking$gameday
```

```
) %>%
  time_length(unit = "years")

# Merge with player info
player_age_offense <-  left_join(
  player_game_birthdate_gameday_offense,
  nfl_players %>% select(-birth_date, -season),
  by = c("player_id" = "gsis_id"))

player_age_defense <-  left_join(
  player_game_birthdate_gameday_defense,
  nfl_players %>% select(-birth_date, -season),
  by = c("player_id" = "gsis_id"))

player_age_kicking <-  left_join(
  player_game_birthdate_gameday_kicking,
  nfl_players %>% select(-birth_date, -season),
  by = c("player_id" = "gsis_id"))

# Add game_id to weekly stats to facilitate merging
nfl_actualStats_game_offense_weekly <- nfl_actualStats_offense_weekly %>%
  left_join(
    player_age_offense[,c("season","week","player_id","game_id")],
    by = c("season","week","player_id"))

nfl_actualStats_game_defense_weekly <- nfl_actualStats_defense_weekly %>%
  left_join(
    player_age_offense[,c("season","week","player_id","game_id")],
    by = c("season","week","player_id"))

nfl_actualStats_game_kicking_weekly <- nfl_actualStats_kicking_weekly %>%
  left_join(
    player_age_offense[,c("season","week","player_id","game_id")],
    by = c("season","week","player_id"))

# Merge with player weekly stats
player_age_stats_offense <- left_join(
  player_age_offense %>% select(-position, -position_group),
  nfl_actualStats_game_offense_weekly,
  by = c(c("season","week","player_id","game_id")))

player_age_stats_defense <- left_join(
  player_age_defense %>% select(-position, -position_group),
  nfl_actualStats_game_defense_weekly,
```

```
  by = c(c("season","week","player_id","game_id")))

player_age_stats_kicking <- left_join(
  player_age_kicking %>% select(-position, -position_group),
  nfl_actualStats_game_kicking_weekly,
  by = c(c("season","week","player_id","game_id")))

player_age_stats_offense$years_of_experience <- as.integer(player_age_stats_offense$years_of_exper
player_age_stats_defense$years_of_experience <- as.integer(player_age_stats_defense$years_of_exper
player_age_stats_kicking$years_of_experience <- as.integer(player_age_stats_kicking$years_of_exper

# Merge player info with seasonal stats
player_seasonal_offense <- left_join(
  nfl_actualStats_offense_seasonal,
  nfl_players %>% select(-position, -position_group, -season),
  by = c("player_id" = "gsis_id")
)

player_seasonal_defense <- left_join(
  nfl_actualStats_defense_seasonal,
  nfl_players %>% select(-position, -position_group, -season),
  by = c("player_id" = "gsis_id")
)

player_seasonal_kicking <- left_join(
  nfl_actualStats_kicking_seasonal,
  nfl_players %>% select(-position, -position_group, -season),
  by = c("player_id" = "gsis_id")
)

# Calculate age
season_startdate <- nfl_schedules %>%
  group_by(season) %>%
  summarise(startdate = min(gameday, na.rm = TRUE))

player_seasonal_offense <- player_seasonal_offense %>%
  left_join(
    season_startdate,
    by = "season"
  )

player_seasonal_defense <- player_seasonal_defense %>%
  left_join(
    season_startdate,
```

```
    by = "season"
  )

player_seasonal_kicking <- player_seasonal_kicking %>%
  left_join(
    season_startdate,
    by = "season"
  )

player_seasonal_offense$age <- interval(
  start = player_seasonal_offense$birth_date,
  end = player_seasonal_offense$startdate
) %>%
  time_length(unit = "years")

player_seasonal_defense$age <- interval(
  start = player_seasonal_defense$birth_date,
  end = player_seasonal_defense$startdate
) %>%
  time_length(unit = "years")

player_seasonal_kicking$age <- interval(
  start = player_seasonal_kicking$birth_date,
  end = player_seasonal_kicking$startdate
) %>%
  time_length(unit = "years")
```

3.22 Plotting

3.22.1 Rushing Yards per Carry By Player Age

```
# Prepare Data
rushing_attempts <- nfl_pbp %>%
  dplyr::filter(
    season_type == "REG") %>%
    filter(
      rush == 1,
      rush_attempt == 1,
      qb_scramble == 0,
      qb_dropback == 0,
```

```
      !is.na(rushing_yards))

rb_yardsPerCarry <- rushing_attempts %>%
  group_by(rusher_id, season) %>%
  summarise(
    ypc = mean(rushing_yards, na.rm = TRUE),
    rush_attempts = n(),
    .groups = "drop") %>%
  ungroup() %>%
  left_join(
    nfl_players %>% select(-season),
    by = c("rusher_id" = "gsis_id")
  ) %>%
  filter(
    position_group == "RB",
    rush_attempts >= 50) %>%
  left_join(
    season_startdate,
    by = "season"
  )

rb_yardsPerCarry$age <- interval(
  start = rb_yardsPerCarry$birth_date,
  end = rb_yardsPerCarry$startdate
) %>%
  time_length(unit = "years")

# Create Plot
ggplot2::ggplot(
  data = rb_yardsPerCarry,
  ggplot2::aes(
    x = age,
    y = ypc)) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth() +
  ggplot2::labs(
    x = "Rushing Back Age (years)",
    y = "Rushing Yards per Carry/season",
    title = "2023 NFL Rushing Yards Per Carry per Season by Player Age",
    subtitle = "(minimum 50 rushing attempts)"
  ) +
  ggplot2::theme_classic()
```

Figure 3.2 2023 NFL Rushing Yards Per Carry per Season by Player Age

```
# Subset Data
rb_seasonal <- player_seasonal_offense %>%
  filter(position_group == "RB")

# Create Plot
ggplot2::ggplot(
  data = rb_seasonal,
  ggplot2::aes(
    x = age,
    y = rushing_epa)) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth() +
  ggplot2::labs(
    x = "Rushing Back Age (years)",
    y = "Rushing EPA/season",
    title = "2023 NFL Rushing Expected Points Added (EPA) per Season by Player Age"
  ) +
  ggplot2::theme_classic()
```

Figure 3.3 2023 NFL Rushing Expected Points Added (EPA) per Season by Player Age

3.22.2 Defensive and Offensive EPA per Play

Expected points added (EPA) per play by the team with possession.

```
pbp_regularSeason <- nfl_pbp %>%
  dplyr::filter(
    season == 2023,
    season_type == "REG") %>%
  dplyr::filter(!is.na(posteam) & (rush == 1 | pass == 1))

epa_offense <- pbp_regularSeason %>%
  dplyr::group_by(team = posteam) %>%
  dplyr::summarise(off_epa = mean(epa, na.rm = TRUE))

epa_defense <- pbp_regularSeason %>%
  dplyr::group_by(team = defteam) %>%
  dplyr::summarise(def_epa = mean(epa, na.rm = TRUE))

epa_combined <- epa_offense %>%
  dplyr::inner_join(epa_defense, by = "team")

ggplot2::ggplot(
```

```
  data = epa_combined,
  ggplot2::aes(
    x = off_epa,
    y = def_epa)) +
  nflplotR::geom_mean_lines(
    ggplot2::aes(
      x0 = off_epa ,
      y0 = def_epa)) +
  nflplotR::geom_nfl_logos(
    ggplot2::aes(
      team_abbr = team),
      width = 0.065,
      alpha = 0.7) +
  ggplot2::labs(
    x = "Offense EPA/play",
    y = "Defense EPA/play",
    title = "2023 NFL Offensive and Defensive EPA per Play"
  ) +
  ggplot2::theme_classic() +
  ggplot2::scale_y_reverse()
```

Figure 3.4 2023 NFL Offensive and Defensive EPA per Play

4

Player Evaluation

4.1 Getting Started

4.1.1 Load Packages

```
library("tidyverse")
```

4.2 Overview

Evaluating players for fantasy football could be thought of as similar to the process of evaluating companies when picking stocks to buy. You want to evaluate and compare various assets so that you get the assets with the best value.

There are various domains of criteria we can consider when evaluating a football player's fantasy prospects. Potential domains to consider include:

- athletic profile
- historical performance
- health
- age and career stage
- situational factors
- matchups
- cognitive and motivational factors
- fantasy value

The discussion that follows is based on my and others' *impressions* of some of the characteristics that may be valuable to consider when evaluating players. However, the extent to which any factor is actually relevant for predicting

future performance is an empirical question and should be evaluated empirically.

4.3 Athletic Profile

Factors related to a player's athletic profile include factors such as:

- body shape
  - height
  - weight
  - hand size
  - wing span (arm length)
- body function
  - agility
  - strength
  - speed
  - acceleration/explosiveness
  - jumping ability

In terms of body shape, we might consider a player's height, weight, hand size, and wing span (arm length). Height allows players to see over opponents and to reach balls higher in the air. Thus, greater height is particularly valuable for Quarterbacks and Wide Receivers. Heavier players are tougher to budge and to tackle. Greater weight is particularly valuable for Linemen, Fullbacks, and Tight Ends, but it can also be valuable—to a deree—for Quarterbacks, Running Backs, and Wide Receivers. Hand size and wing span is particularly valuable for people catching the ball; thus, a larger hand size and longer wing span are particularly valuable for Wide Receivers and Tight Ends.

In terms of body function, we can consider a player's agility, strength, speed, acceleration/explosiveness, and jumping ability. For Wide Receivers, speed, explosiveness, and jumping ability are particularly valuable. For Running Backs, agility, strength, speed, and explosiveness are particularly valuable.

Many aspects of a player's athletic profile, including tests of speed (40-yard dash), strength (bench press), agility (2-yard shuttle run; three cone drill), and jumping ability (vertical jump; broad jump) are available from the National Football League (NFL) Combine, which is especially relevant for evaluating rookies. We demonstrate how to import data from the NFL Combine in Section 3.5.6. There are also calculators that integrate information about body shape and information from the NFL Combine to determine a player's relative athletic score (RAS) for their position: https://ras.football/ras-calculator/

4.4 Skill

When scouting players, scouts consider not only the player's athletic profile, but also their position-relevant skill. For instance, how good are they are reading the defense, passing the ball, running routes, catching balls, making defenders miss tackles, taking care of the ball, consistency, etc. Scouting and evaluating skill is a complicated endeavor, and even the professional scouts frequently make mistakes in their evaluations and predictions. You can certainly read skill evaluations about various players; however, unlike metrics of athletic profile, we do not have direct access to the player's underlying skill. Some may say, "You know it when you see it." But, this is not particularly useful when trying to identify players who are undervalued or overvalued—because the skill evaluations are likely already "baked into" a player's projections. Because we do not have direct access to a player's skill, we tend to rely on indirect metrics of their ability, such as historical performance.

4.5 Historical Performance

4.5.1 Overview

> "The best predictor of future behavior is past behavior." – Unknown

> "Past performance does not guarantee future results." – A common disclaimer about investments.

Factors relating to historical performance to consider could include:

- performance in college

  - draft position

- performance in the NFL
- efficiency
- consistency

Compared to tests of speed, power, and agility at the NFL Combine, collegiate performance is a stronger predictor of performance in the NFL (Lyons et al., 2011). That is, previous sports performance is the best predictor of future performance (for a review, see Den Hartigh et al., 2018). Thus, it is important to consider a player's past performance. However, the extent to which historical performance may predict future performance may depend on many factors such as (a) the similarity of the prior situation to the current situation, (b) how long ago the prior situation was, and (c) the extent to which the player (or situation) has changed in the interim. For rookies, the player does not have prior seasons of performance in the NFL to draw upon. Thus, when evaluating rookies, it can be helpful to consider their performance in college or in their prior leagues. However, there are large differences between the situation in college and the situation in the NFL, so prior success in college may not portend future success in the NFL. An indicator that intends to be prognostic of future performance, and that accounts for past performance, is a player's draft position—that is, how early (or late) was a player selected in the NFL Draft. The earlier a player was selected in the NFL Draft, the greater likelihood that the player will perform well; however, this is somewhat countered by the fact that the teams with the highest draft picks tend to be the worst based on the prior season's record.

For players who have played in the NFL, past performance becomes more relevant because, presumably, the prior situation is more similar (than was their situation in college) to their current situation. Nevertheless, lots of things change from game to game and season to season: injuries, coaches, coaching strategies, teammates, etc. So just because a player performed well or poorly in a given game or season does not necessarily mean that they will perform similarly in subsequent games/seasons. Nevertheless, historical performance is one of the best indicators we have.

We demonstrate how to import historical player statistics in Section 3.5.12. We demonstrate how to calculate historical player statistics in Section 3.21.1. We demonstrate how to calculate historical fantasy points in Section 3.21.2.

4.5.2 Efficiency

In addition to how many fantasy points a player scores in terms of historical performance, we also care about efficiency and consistency. How efficient were they given the number of opportunities they had? If they were relatively more

efficient, they will likely score more points than many of their peers when given more opportunities. If they were relativelly inefficient, their capacity to score fantasy points may be more dependent on touches/opportunities. Efficiency might be operationalized by indicators such as yards per passing attempt, yards per rushing attempt, yards per target, yards per reception, etc.

4.5.3 Consistency

In terms of consistency, how consistent was the player they from game to game and from season to season? For instance, we could examine the standard deviations of players' fantasy points across games in a given season. However, the standard deviation tends to be upwardly biased as the mean increases. So, we can account for the player's mean fantasy points per game by dividing their game-to-game standard deviation of fantasy points (σ) by their mean fantasy points across games (μ). This is known as the coefficient of variation (CV), which is provided in Equation 4.1.

$$CV = \frac{\sigma}{\mu} \tag{4.1}$$

Players with a lower standard deviation and a lower coefficient of variation (of fantasy points across games) are more consistent. In the example below, Player 2 might be preferable to Player 1 because Player 2 is more consistent; Player 1 is more "boom-or-bust." Despite showing a similar mean of fantasy points across weeks, Player 2 shows a smaller week-to-week standard deviation and coefficient of variation.

```
set.seed(1)

playerScoresByWeek <- data.frame(
  player1_scores = rnorm(17, mean = 20, sd = 7),
  player2_scores = rnorm(17, mean = 20, sd = 4),
  player3_scores = rnorm(17, mean = 10, sd = 4),
  player4_scores = rnorm(17, mean = 10, sd = 1)
)

consistencyData <- data.frame(t(playerScoresByWeek))

weekNames <- paste("week", 1:17, sep = "")

names(consistencyData) <- weekNames
row.names(consistencyData) <- NULL

consistencyData$mean <- rowMeans(consistencyData[,weekNames])
consistencyData$sd <- apply(consistencyData, 1, sd)
```

```r
consistencyData$cv <- consistencyData$sd / consistencyData$mean

consistencyData$player <- c(1, 2, 3, 4)

consistencyData <- consistencyData %>%
  select(player, mean, sd, cv, week1:week17)

round(consistencyData, 2)
```

```
  player  mean   sd   cv week1 week2 week3 week4 week5 week6 week7 week8 week9
1      1 20.60 6.47 0.31 15.61 21.29 14.15 31.17 22.31 14.26 23.41 25.17 24.03
2      2 20.61 3.35 0.16 23.78 23.28 22.38 23.68 23.13 20.30 12.04 22.48 19.78
3      3 10.32 2.65 0.26  4.49  8.34  8.42  9.76 14.40 13.05  9.34  8.99 12.79
4      4 10.19 1.11 0.11  9.39 10.34  8.87 11.43 11.98  9.63  8.96 10.57  9.86
  week10 week11 week12 week13 week14 week15 week16 week17
1  17.86  30.58  22.73  15.65   4.50  27.87  19.69  19.89
2  19.38  14.12  18.09  21.67  25.43  19.59  21.55  19.78
3  12.23   7.24   7.17  11.46  13.07   9.55  13.52  11.59
4  12.40   9.96  10.69  10.03   9.26  10.19   8.20  11.47
```

4.6 Health

Health-related factors to consider include:

- current injury status
- injury history

It is also important to consider a player's past and current health status. In terms of a player's current health status, it is important to consider whether they are injured or are playing at less than 100% of their typical health. In terms of a player's prior health status, one can consider their injury history, including the frequency and severity of injuries and their prognosis.

We demonstrate how to import injury reports in Section 3.5.13.

4.7 Age and Career Stage

Age and career stage-related factors include:

- age
- experience
- touches

A player's age is relevant because of important age-related changes in a player's speed, ability to recover from injury, etc. A player's experience is relevant because players develop knowledge and skills with greater experience. A player's prior touches/usage is also relevant, because it speaks to how many hits a player may have taken. For players who take more hits, it may be more likely that their bodies "break down" sooner.

4.8 Situational Factors

Situational factors one could consider include:

- team quality
- role on team
- teammates
- opportunity and usage
  - snap count
  - touches/targets
  - red zone usage

Football is a team sport. A player is embedded within a broader team context; it is important to consider the strength of their team context insofar as it may support— or detract from—a player's performance. For instance, for a Quarterback, it is important to consider how strong the pass blocking is from the Offensive Line. Will they have enough time to throw the ball, or will they be constantly under pressure to be sacked? It is also important to consider the strength of the pass catchers—the Wide Receivers and Tight Ends. For a Running Back, it is important to consider how strong the run blocking is from the Offensive Line. For a Wide Receiver, it is important to consider how strong the pass blocking is, and how strong the Quarterback is.

It is also important to consider a player's role on the team. Is the player a starter or a backup? Related to this, it is important to consider the strength of one's teammates. For a given Running Back, if a teammate is better at running the ball, this may take away from how much the player sees the field. For a given Wide Receiver, if a teammate is better at catching the ball, this may take some targets away from the player. However, the team's top

defensive back is often matched up against the team's top Wide Receiver. So, if the team's top Wide Receiver is matched up against a particularly strong Defensive Back, the second- and third-best Wide Receivers may more targets than usual.

It is also important to consider a player's opportunity and usage, which are influenced by many factors, including the skill of the player, the skill of their teammates, the role of the player on the team, the coaching style, the strategy of the opposing team, game scripts, etc. In terms of the player's opportunity and usage, how many snaps do they get? How many touches and/or targets do they receive? Being on the field for more snaps and receiving more touches and/or targets means that the player has more opportunities to score fantasy points. Are they targeted in the red zone? Red zone targets are more likely to lead to touchdown scoring opportunities, which are particularly valuable in fantasy football.

4.9 Matchups

Matchup-related factors to consider include:

- strength of schedule
- weekly matchup

Another aspect to consider is how challenging their matchup(s) and strength of schedule is. For a Quarterback, it is valuable to consider how strong the oppenent's passing defense is. For a Running Back, how strong is the running defense? For a Wide Receiver, how strong is the passing defense and the Defensive Back that is likely to be assigned to guard them?

4.10 Cognitive and Motivational Factors

Other factors to consider include cognitive and motivational factors. Some coaches refer to these as the "X Factor" or "the intangibles." However, just as any other construct in psychology, we can devise ways to operationalize them. Insofar as they are observable, they are measurable.

Cognitive and motivational factors one could consider include:

- reaction time
- knowledge and intelligence
- work ethic and mental toughness
- incentives
  - contract performance incentives
  - whether they are in a contract year

A player's knowledge, intelligence, and reaction time can help them gain an upper-hand even when they may not be the fastest or strongest. A player's work ethic and mental toughness may help them be resilient and persevere in the face of challenges. Contact-related incentives may lead a player to put forth greater effort. For instance, a contract may have a performance incentive that provides a player greater compensation if they achieve a particular performance milestone (e.g., receiving yards). Another potential incentive is if a player is in what is called their "contract year" (i.e., the last year of their current contract). If a player is in the last year of their current contract, they have an incentive to perform well so they can get re-signed to a new contract.

4.11 Fantasy Value

4.11.1 Sources From Which to Evaluate Fantasy Value

There are several sources that one can draw upon to evaluate a player's fantasy value:

- expert or aggregated rankings
- layperson rankings
  - players' Average Draft Position (ADP) in other league snake drafts
  - players' Average Auction Value (AAV) in other league auction drafts
- expert or aggregated projections

4.11.1.1 Expert Fantasy Rankings

Fantasy rankings (by so-called "experts") are provided by many sources. To reduce some of the bias due to a given source, some services aggregate projections across sources, consistent with a "wisdom of the crowd" approach. FantasyPros[1] aggregates fantasy rankings across sources. Fantasy Football Ana-

[1] https://www.fantasypros.com/nfl/rankings/consensus-cheatsheets.php

lytics[2] creates fantasy rankings from projections that are aggregated across sources (see the webapp here: https://apps.fantasyfootballanalytics.net).

4.11.1.2 Layperson Fantasy Rankings: ADP and AAV

Average Draft Position (ADP) and Average Auction Value (AAV), are based on league drafts, mostly composed of everyday people. ADP is based on snake drafts, whereas AAV is based on auction drafts. Thus, ADP and AAV are consistent with a "wisdom of the crowd" approach, and I refer to them as forms of rankings by laypeople. ADP data are provided by FantasyPros[3]. AAV data are also provided by FantasyPros[4].

4.11.1.3 Projections

Projections are provided by various sources. Projections (and rankings, for that matter) are a bit of a black box. It is often unclear how they were derived by a particular source. That is, it is unclear how much of the projection was based on statistical analysis versus conjecture.

To reduce some of the bias due to a given source, some services aggregate projections across sources, consistent with a "wisdom of the crowd" approach. Projections that are aggregated across sources are provided by Fantasy Football Analytics[5] (see the webapp here: https://apps.fantasyfootballanalytics.net) and by FantasyPros[6].

4.11.1.4 Benefits of Using Projections Rather than Rankings

It is important to keep in mind that rankings, ADP, and AAV are specific to roster and scoring settings of a particular league. For instance, in point-per-reception (PPR) leagues, players who catch lots of passes (Wide Receivers, Tight Ends, and some Running Backs) are valued more highly. As another example, Quarterbacks are valued more highly in 2-Quarterback leagues. Thus, if using rankings, ADP, or AAV, it is important to find ones from leagues that mirror—as closely as possible—your league settings.

Projected statistics (e.g., projected passing touchdowns) are agnostic to league settings and can thus be used to generate league-specific fantasy projections and rankings. Thus, projected statisitics may be more useful than rankings because they can be used to generate rankings for your particular league settings. For instance, if you know how many touchdowns, yards, and interceptions a

[2] https://fantasyfootballanalytics.net
[3] https://www.fantasypros.com/nfl/adp/overall.php
[4] https://www.fantasypros.com/nfl/auction-values/calculator.php
[5] https://fantasyfootballanalytics.net
[6] https://www.fantasypros.com/nfl/auction-values/calculator.php

Quarterback is a projected to throw (in addition to any other relevant categories for the player, e.g., rushing yards and touchdowns), you can calculate how many fantasy points the Quarterback is expected to gain in your league (or in any league). Thus, you can calculate ranking from projections, but you cannot reverse engineer projections from rankings.

4.11.2 Indices to Evaluate Fantasy Value

Based on the sources above (rankings, ADP, AAV, and projections), we can derive multiple indices to evaluate fantasy value. There are many potential indices that can be worthwhile to consider, including a player's:

- dropoff
- value over replacement player (VORP)
- uncertainty

4.11.2.1 Dropoff

A player's *dropoff* is the difference between (a) the player's projected points and (b) the projected points of the next-best player at that position.

4.11.2.2 Value Over Replacement Player

Because players from some positions (e.g., Quarterbacks) tend to score more points than players from other positions (e.g., Wide Receivers), it would be inadvisable to compare players across different positions based on projected points. In order to more fairly compare players across positions, we can consider a player's value over a typical replacement player at that position (shortened to "value over replacement player"). A player's *value over a replacement player* (VORP) is the difference between (a) a player's projected fantasy points and (b) the fantasy points that you would be expected to get from a typical bench player at that position. Thus, VORP provides an index of how much added value a player provides.

4.11.2.3 Uncertainty

A player's *uncertainty* is how much variability there is in projections or rankings for a given player across sources. For instance, consider a scenario where three experts provide ratings about two players, Player A and Player B. Player A is projected to score 300, 310, and 290 points by experts 1, 2, and 3, respectively. Player B is projected to score 400, 300, and 200 points by experts 1, 2, and 3, respectively. In this case, both players are (on average) projected to score the same number of points (300).

```r
exampleData <- data.frame(
  player = c(rep("A", 3), rep("B", 3)),
  expert = c(1:3, 1:3),
  projectedPoints = c(300, 310, 290, 400, 300, 200)
)

playerA_mean <- mean(exampleData$projectedPoints[which(exampleData$player == "A")])
playerB_mean <- mean(exampleData$projectedPoints[which(exampleData$player == "B")])

playerA_sd <- sd(exampleData$projectedPoints[which(exampleData$player == "A")])
playerB_sd <- sd(exampleData$projectedPoints[which(exampleData$player == "B")])

playerA_cv <- playerA_mean / playerA_sd
playerB_cv <- playerB_mean / playerB_sd
```

```r
playerA_mean
```

```
[1] 300
```

```r
playerB_mean
```

```
[1] 300
```

However, the players differ considerably in their uncertainty (i.e., the source-to-source variability in their projections), as operationalized with the standard deviation and coefficient variation of projected points across sources for a given player.

```r
playerA_sd
```

```
[1] 10
```

```r
playerB_sd
```

```
[1] 100
```

```r
playerA_cv
```

```
[1] 30
```

```
playerB_cv
```

```
[1] 3
```

Here is a depiction of a density plot of projected points for a player with a low, medium, and high uncertainty:

```
playerA <- rnorm(1000000, mean = 150, sd = 5)
playerB <- rnorm(1000000, mean = 150, sd = 15)
playerC <- rnorm(1000000, mean = 150, sd = 30)

mydata <- data.frame(playerA, playerB, playerC)

mydata_long <- mydata %>%
  pivot_longer(
    cols = everything(),
    names_to = "player",
    values_to = "points"
  ) %>%
  mutate(
    name = case_match(
      player,
      "playerA" ~ "Player A",
      "playerB" ~ "Player B",
      "playerC" ~ "Player C",
    )
  )

ggplot2::ggplot(
  data = mydata_long,
  ggplot2::aes(
    x = points,
    fill = name
  )
) +
  ggplot2::geom_density(alpha = .3) +
  ggplot2::labs(
    x = "Players' Projected Points",
    title = "Density Plot of Projected Points for Three Players"
  ) +
  ggplot2::theme_classic() +
  ggplot2::theme(legend.title = element_blank())
```

Figure 4.1 Density Plot of Projected Points for Three Players

Uncertainty is not necessarily a bad characteristic of a player's projected points. It just means we have less confidence about how the player may be expected to perform. Thus, players with greater uncertainty are risky and tend to have a higher upside (or ceiling) and a lower downside (or floor).

4.12 Putting it Altogether

After performing an evaluation of the relevant domain(s) for a given player, then one must integrate the evaluation information across domains to make a judgment about a player's overall value. When thinking about a player's value, it can be worth thinking of a player's upside and a player's downside. Player that are more consistent may show higher downside but a lower upside. Younger, less experienced players may show a higher upside but a lower downside.

The extent to which you prioritize a higher upside versus a higher downside may depend on many factors. For instance, when drafting players, you may prioritize drafting players with the highest downside (i.e., the safest players), whereas you may draft sleepers (i.e., players with higher upside) for your bench. When choosing which players to start in a given week, if you are predicted to beat a team handily, it may make sense to start the players with

the highest downside. By contrast, if you are predicted to lose to a team by a good margin, it may make sense to start the players with the highest upside.

5

The Fantasy Draft

5.1 Getting Started

5.1.1 Load Packages

5.2 Types of Fantasy Drafts

There are several types of drafts in fantasy football. The most common types of drafts are snake drafts and auction drafts.

5.2.1 Snake Draft

In a snake draft, the participants (i.e., managers) are assigned a draft order. In the first round, the managers draft in that order. In the second round, the managers draft in reverse order. It continues to "snake" in this way, round after round, so that the person who has the first pick in a given round has the last pick in the next round, and whoever has the last pick in a given round has the first pick in the next round.

5.2.2 Auction Draft

In an auction draft, the managers are assigned a nomination order and there is a salary cap (e.g., $200). The first manager chooses which player to nominate. Then, the managers bid on that player like in an auction. In order to bid, the manager must raise the price by at least $1. If two managers want to obtain the same player, they may continue to raise the amount until one manager backs out and is no longer to bid by raising the price. The highest bidder wins (i.e., drafts) that player. Then, the second manager nominates a player, and the managers bid on that player. This process repeats until all teams have drafted their allotment of players.

5.2.3 Comparison

Snake drafts are more common than auction drafts. Snake drafts tend to be quicker than auction drafts. However, auction drafts are more fair than snake drafts. In an auction draft, unlike a snake draft, all players are available to all teams. For instance, in a snake draft, the first 9 players drafted are unavailable to the 10th pick of the first round. So, if you have the 10th pick and want the top-ranked player, this player would not be available to you in the snake draft. However, in the auction draft, every player is available to every manager, so long as the manager is able and willing to bid enough.

5.3 Draft Strategy

5.3.1 Overview

There is no one "right" draft strategy. As noted by Lee & Liu (2022) in their analysis of fantasy drafts, the effectiveness of any draft strategy depends on the strategies of the other managers in the league. Sometimes it works best to "zig" when everyone else is "zagging". For instance, if you notice that everyone else is drafting Wide Receivers, this may mean that other managers are over-valuing Wide Receivers, and this could be a nice opportunity to draft a Running Back for good value.

In general, you will first want to generate the rankings you will use to select which players to prioritize. You may generate your rankings based one or more of the following:

- your evaluation of players[1]
- expert or aggregated rankings
- layperson rankings
  - players' Average Draft Position (ADP) in other league drafts (for snake drafts)
  - players' Average Auction Value (AAV) in other league drafts (for auction drafts[2])
- expert or aggregated projections
- indices derived from rankings and projections

[1] player-evaluation.qmd
[2] sec-draftStrategyAuction

Section 4.11.1 describes where to obtain aggregated rankings, aggregated projections, ADP, and AAV data.

An important concept in the draft is "dropoff", which is described in Section 4.11.2.1. *Dropoff* at a given position, is the difference—in terms of projected fantasy points—between (a) the best available player remaining at that position and (b) the second-best available player remaining at that position. If there is a bigger dropoff at a given position, there may be greater value in drafting the top player from that position. For instance, consider the following scenario: "Quarterback A" is projected to score 325 points, and "Quarterback B" is projected to score 320 points. "Tight End A" is projected to score 230 points, and "Tight End B" is projected to score 150 points. In this example, there is a much greater dropoff for Tight Ends than there is for Quarterbacks. Thus, even though "Quarterback A" is projected to score more points than "Tight End A", "Tight End A" may be more valuable because there is still a good Quarterback available if someone else drafts "Quarterback A".

Another important concept is a player's value over a typical replacement player at that position (shortened to "value over replacement player"; VORP), which is described in Section 4.11.2.2.

Another important concept is a player's uncertainty, which is described in Section 4.11.2.3.

In both snake and auction draft formats, your goal is to draft the team whose weekly starting lineup scores the most points and thus the collection of players with the greatest VORP. For your starting lineup, it may make sense—especially with your earliest selections—when comparing two players with equivalent VORP, to prioritize players with higher consistency and lower uncertainty, because they may be considered "safer" with a higher floor. However, when drafting players for your bench, it make make more sense to prioritize high-risk, high reward players with greater uncertainty, because they may have a higher ceiling. Players with a higher ceiling have a potential to be "sleepers"—players who are valued low (i.e., with a high ADP or low AAV) and who outperform their valuation. Note that, although players with greater uncertainty are high-risk, high-reward players, selecting this kind of a player for your bench (i.e., in a late round or for a small cost) is a *lower* risk selection, because you have less to lose with later/lower-cost picks. That is, even though the *player* is higher risk, selecting a higher risk player for your bench is a lower risk *decision*.

The Spurs in the National Basketball Association (NBA) were well-reputed for excelling in this draft strategy[3] (archived at https://perma.cc/X7NW-WZC6). They frequently used their second-round picks to draft high-risk, high-reward players. Sometimes, the secound round pick was a bust, but they

[3] https://harvardsportsanalysis.org/2013/11/beating-the-nba-draft-does-any-team-outperform-expectations/

have little to lose with a failed second round pick. Other times, their second round picks—including Willie Anderson, DeJuan Blair, Goran Dragic, Luis Scola, and Manu Ginóbili—greatly outperformed expectations. Thanks, in part, to this draft strategy, the team showed strong extended success for nearly three decades from 1989 through the late-2010s.

However, the draft strategies to achieve the "optimal lineup" differ between snake versus auction drafts.

5.3.2 Snake Draft

In general, your goal is to draft the team whose weekly starting lineup has the greatest VORP. Consequently, you are often looking to pick the player with the highest VORP at a given selection, while keeping in mind (a) the dropoff of players at other positions and (b) which players may be available at subsequent picks so that you do not sacrifice too much later value with a given selection. For instance, if a particular Quarterback has a slightly higher VORP than a particular Running Back, but the Quarterback is likely to be available at the manager's next pick but the Running Back is likely to be unavailable at their next pick, it might make more sense to draft the Running Back.

5.3.3 Auction Draft

According to an analysis[4] by the Harvard Sports Analysis Collective (archived at https://perma.cc/P7RX-92UU), the majority of the manager's salary cap should be spent on the starting lineup, and you should spend less on bench players. This is known as the "stars and scrubs" draft strategy. Based on the analysis, the author recommended applying a 10% premium to the top players and a 10% discount to the lower-tiered players. The idea behind the approach is that a player on your bench does not contribute to the team's points and, thus, most players drafted to your bench do not contribute much to the team's points throughout the season. That said, bench players can be important in the case of a starter's injury or under-performance. So, it is recommended to draft starters with lower uncertainty who are safer. In contrast to your starting lineup, you may look to draft players on your bench who have greater uncertainty for their high reward potential in a low-risk selection given the lower price.

An alternative to the "stars and scrubs" approach is to wait to draft more "high-value" players after other managers have over-paid for players.

[4] https://harvardsportsanalysis.wordpress.com/wp-content/uploads/2012/04/fantasyfootballdraftanalysis1.pdf

6

Research Methods

6.1 Getting Started

6.1.1 Load Packages

6.2 Sample vs Population

In research, it is important to distinguish between the sample and the target population. The target *population* is who you want your study's findings to generalize to. For instance, if we want our findings to lead to inferences we can draw regarding all current NFL players, then NFL players are our target population. However, despite our best efforts to recruit all NFL players into our study, we may not succeed in doing that. The participants (i.e., people or players) who we successfully recruit to be in our study represent our *sample*. The number of participants in the study is our *sample size*.

It is rare for the sample to include all people who are in the target population. It can be costly to recruit large samples, and many potential participants may decline to participate for a variety of reasons (insufficient time, lack of interest in the study, distrust of scientists, etc.). Thus, our goals are (a) to recruit as many people from the population as possible and (b) for the sample to be as *representative* of the population as possible.

For increasing the representativeness of the sample (with respect to the population), we might conduct a *random sample*, in which each person in the population (i.e., each NFL player) has equal likelihood of being selected. For instance, we might randomly select 250 players to recruit to the study. True random samples, though strong in aspiration, are difficult and costly to achieve. In reality, many researchers conduct convenience sampling. A convenience sample is recruited because it is convenient (i.e., less costly and time-consuming).

For instance, many studies examine college students—in part, because they are easy to recruit. If our target population is NFL players but we are unable

to recruit NFL players into our study, we could easily recruit a large sample of college students. Although the convenience sample may afford a very large sample, the college student sample may not be representative of the target population (NFL players). Thus, the findings in our study may not *generalize* to NFL players—that is, what we learn in college students may not apply in the same way among NFL players. For instance, if we learn that consumption of sports drinks (compared to drinking only water) improves running speed among college students, that may not be the case among NFL players.

6.3 Research Designs

There are three broad types of research designs:

- experiment
- correlational/observational study
- case study

6.3.1 Experiment

In an *experiment*, there are one or more things (i.e., variables) that we manipulate to see how the manipulation influences the process of interest. The variable that we manipulate is the *independent variable*. By contast, the *dependent variable* is the variable that we evaluate to determine whether it was influenced by the manipulation (i.e., by the independent variable).Besides the independent and dependent variables, the researcher attempts to hold everything else constant through processes including standardization and random assignment. *Standardization* involves using the same procedures to assess each participant, so that scores can be fairly compared across participants (and groups). Random assignment involves randomly assigning participants to conditions of the independent variable, so the people in each condition are comparable and do not differ systematically.

For instance, we may be interested to evaluate whether players perform better (e.g., run faster) if they drink a sports drink compared when they drink only water. Our hypothesis might be that players will be expected to perform better when they drink a sports drink (compared to when they drink only water). To this this research question and hypothesis, we might conduct an experiment by randomly assigning some players during practice to receive a sports drink and some players to receive only water. In this case, our independent variable is whether the player receives a sports drink. Our dependent variable might be their 40-yard dash time during practice.

6.3.2 Correlational/Observational Study

In a correlational (aka observational) study, we do not manipulate a variable to see how the manipulation influences another variable. Instead, we examine how two variables, a predictor and an outcome variable, are associated. The hypothesized cause is called the predictor variable. The hypothesized effect is called the outcome variable. In this way, the predictor variable is similar to the independent variable, and the outcome variable is similar to the dependent variable. However, unlike the independent and dependent variables in an experiment, the predictor and outcome variables in a correlational study are not manipulated.

For instance, to use a correlational study to test the possibility that players who drink sports drinks perform better than players who drink only water, we could examine whether the players who drink sports drinks during a game score more fantasy points than players who drink only water during the game. In this case, our predictor variable is whether the players drinks sports drinks during a game. Our outcome variable is the number of fantasy points the player scored.

6.3.2.1 Correlation Does Not Imply Causation

As the maxim goes, "correlation does not imply causation"—just because two variables are associated does not necessarily mean that they are causally related.

Just because `X` is associated with `Y` does not mean that `X` causes `Y`. Consider that you find an association between variables `X` and `Y`.

There are several reasons why you might observe an association between `X` and `Y`:

- `X` causes `Y`
- `Y` causes `X`
- `X` and `Y` are bidirectional: `X` causes `Y` and `Y` causes `X`
- a third variable (i.e., confound), `Z`, influences both `X` and `Y`
- the association between `X` and `Y` is spurious

For instance, one possibility is that the association we observed reflects our hypothesis that `X` causes `Y`, as depicted in Figure 6.1. That is, consumption of more sports drink may improve players' performance.

However, a second possibility is that the association reflects the opposite direction of effect, where `Y` actually causes `X`, as depicted in Figure 6.2. For instance, greater performance may lead players to drink more sports drink (rather than the reverse).

Figure 6.1 Hypothesized Causal Effect Based on an Observed Association Between X and Y, Such That X Causes Y.

Figure 6.2 Reverse (Opposite) Direction of Effect From the Hypothesized Effect, Where Y Causes X.

A third possibility is that the association reflects a bidirectional effect, where X causes Y and Y causes X, as depicted in Figure 6.3. For instance, consumption of more sports drink may improve players' performance, and greater performance in turn may lead players to drink more sports drink.

Figure 6.3 Bidirectional Effect Between X and Y, such that X causes Y and Y causes X.

A fourth possibility is that the association could reflect the influence of a third variable. If a third variable is a common cause of each and accounts for their association, it is a *confound*. An observed association between X and Y could reflect a confound—i.e., a cause (Z) that influences both X and Y, which explains why X and Y are correlated even though they are not causally related. A third variable confound that is a common cause of both X and Y is depicted in Figure 6.4. For instance, it may not be that sport drink consumption per se influences player performance; rather, it may be that players who are more intelligent or have more financial resources tend to drink more sports drinks and also tend to perform better. In this case, intelligence or financial resources may be a confound that influences both sports drink consumption and player

performance, but sports drink consumptions—though correlated with player performance—does not influence player performance.

For another example, consider that ice cream sales are associated with shark attacks. It is unlikely that more people eating ice creams leads to shark attacks. There is a likely a third variable—heat waves—that is a confound because it influences both ice cream sales and shark attacks and explains their association.

Figure 6.4 Confounded Association Between `X` and `Y` due to a Common Cause, `Z`.

Lastly, the association might be spurious. It might just reflect random variation (i.e., chance), and that when tested on an independent sample, what appeared as an association in the original dataset may not hold when testing the association in a new dataset.

6.3.3 Case Study

In a case study, we assess a small sample of individuals (commonly only one person or a few people), often with rich qualitative information. Themes may be coded from the qualitative information, which may help inform inferences about whether some process may have played a role in influencing the outcome of interest. The inferences are then drawn in a subjective, qualitative way. Testimonials and anecdotes are examples are case studies.

For instance, to use a case study to evalute the possibilty that players who drink sports drinks perform better than players who drink only water, we could conduct an in-depth interview with a player. In the interview, we might ask the player how they performed in games with versus without a sports drink and have them discuss whether they believe the sports drink improved their performance (and if so, how). Then, based on the player's responses, we might code the responses to extract themes and to make a qualitative judgement of whether or not the player likely performed better during games in which they had a sports drink.

6.3.4 Other Features of the Research Design

6.3.4.1 Number of Timepoints

In addition to whether the research design is an experiment, correlational/observational study, or a case study, a research design can also have one or multiple timepoints. The differing number of timepoints allow studies to be characterized as one of the following:

- cross-sectional
- longitudinal

6.3.4.1.1 Cross-Sectional

A *cross-sectional study* is a study with one timepoint.

For instance, in a cross-sectional study evaluating whether having a sports drink improves player performance, we might assess players' drinking behavior and performance during only game 1.

Cross-sectional studies are more common than longitudinal studies because cross-sectional studies are less costly and time-consuming. They can provide a helpful starting point to test findings more rigorously in subsequent longitudinal studies.

6.3.4.1.2 Longitudinal Design

A *longitudinal study* is a study with more than one timepoint. When the same measures are assessed at each of multiple timepoints, we refer to this as a "repeated measures" design.

In a longitudinal study evaluating whether having a sports drink improves player performance, we might assess players' drinking behavior and performance during each game of the season, and possibly across multiple seasons.

Longitudinal studies are less common than cross-sectional studies because longitudinal studies are more costly and time-consuming. Nevertheless, longitudinal studies can allow us test our hypotheses more rigorously, because they can allow us to test whether changes in the predictor/indepdnent variable leads to changes in the outcome/dependent variable. Thus, compared to cross-sectional studies, longitudinal studies can provide greater confidence in causal inferences.

6.3.4.2 Within- or Between-Subject

A research design can also be within-subject, between-subject, or both. A study can involve both within-subject and between-subject comparisons if one predictor/independent variable is within-subject and another predictor/independent variable is between-subject.

6.3.4.2.1 Within-Subject Design

A *within-subject design* is one in which each participant (i.e., person or player) receives multiple levels of the independent variable (or predictor).

For instance, in an experiment evaluating whether having a sports drink improves player performance, we might assign players to drink the sports drink in the first half of the game and to drink only water in the second half of the game. Or we could assign some of the players to drink sports drink in the first half and water in the second half, and assign the other players to drink water in the first half and sports drink in the second half.

In a correlational study evaluating whether having a sports drink improves player performance, we might evaluate how within-person changes in sports drink consumption are associated with within-person changes in performance. That is, we could evaluate, when a given player has a sports drink (or more sports drinks), do they perform better than when the same individual has only water (or fewer sports drinks)?

Within-subject designs tend to have greater statistical power than between-subject designs. However, within-subject designs often have *carryover effects*. For instance, consider the study in which we assign players to drink only water in the first and third quarters and to drink sports drink in the second and fourth quarters (an A-B-A-B design). Drinking sports drink in the second quarter could increase how much hydration a player has throughout the rest

of the game, which could lead to altered performance in the third and fourth quarters that is not due to what they drink in third and fourth quarters.

6.3.4.2.2 Between-Subject Design

A *between-subject design* is one in which each participant (i.e., person or player) receives only one level of the independent variable.

For instance, in an experiment evaluating whether having a sports drink improves player performance, we might assign some players to drink the sports drink but the other players to drink only water.

In a correlational study evaluating whether having a sports drink improves player performance, we might evaluate whether people who drink sports drinks tend to perform better than players who drink only water. Or, we could evaluate whether players who drink more sports drinks perform better than players who drink fewer sports drinks (i.e., whether the number of sports drinks consumed during a game is correlated with player performance).

6.4 Research Design Validity

Research design validity involves the accuracy of inferences from a study. There are three types of research design validity:

- internal validity
- external validity
- conclusion validity

6.4.1 Internal Validity

Internal validity is the extent to which we can be confident that the associations identified in the study are causal.

6.4.2 External Validity

External validity is the extent to which we can be confident that findings from the study play out similarly in the real world—that is, the findings generalize to the target population.

6.4.3 Tradeoffs Between Internal and External Validity

There is a tradeoff between internal and external validity—a single research design cannot have both high internal and high external validity. Each study and design has weaknesses. Some research designs are better suited for making causal inferences, whereas other designs tend to be better suited for making inferences that generalize to the real world. The research design that is best suited to making causal inferences is an experiment because it is the design in which the researcher has the greatest control over the variables. Thus, experiments tend to have higher internal validity than other research designs. However, by manipulating one variable and holding everything else constant, the research takes place in a very standardized fashion that can become like studying a process in a vacuum. So, even if a process is theoretically causal in a vacuum, it may act differently in the real world when it interacts with other processes.

Correlational designs have greater capacity for external validity than experimental designs because the participants can be observed in their natural environments to evaluate how variables are related in the real world. However, the greater external validity comes at a cost of lower internal validity. Correlational designs are not well-positioned to make causal inferences. Correlational studies can account for potential confounds using *covariates* or for the reverse direction of effect using longitudinal designs, but the researcher has less control over the variables than in an experiment.

As the internal validity of a study's design increases, its external validity tends to decrease. The greater control we have over variables (and, therefore, have greater confidence about causal inferences), the lower the likelihood that the findings reflect what happens in the real world because it is studying things in a metaphorical vacuum. Because no single research design can have both high internal and external validity, scientific inquiry needs a combination of many different research designs so we can be more confident in our inferences—experimental designs for making causal inferences and correlational designs for making inferences that are more likely to reflect the real world.

Case studies, because they have smaller sample sizes and inferences drawn in a subjective, qualitative way, tend to have lower external validity than both experimental and correlational studies. Case studies also tend to have lower internal validity because they have less control over variables, and thus fail to remove the possibility of illusory correlations, potential confounds, or the reverse direction of effect. Thus, case studies are among the weakest forms of evidence. Nevertheless, case studies can still be useful for generating hypotheses that can then be tested empirically with a larger sample in experimental or correlational studies.

6.4.4 Conclusion Validity

Conclusion validity is the extent to which a study's conclusions are reasonable about the association among variables based on the data. That is, were the correct statistical analyses performed, and are the interpretations of the findings from those analyses correct?

6.5 Mediation vs Moderation

Both types of effects involve (at least) three variables:

1. An independent/predictor variable, which will be labeled as `X`.
2. A dependent/outcome variable, which will be labeled as `Y`.
3. The mediator or moderator variable, which will be labeled as `M`.

A mnemonic to help remember the difference between mediation and moderation is in Figure 6.5.

6.5.1 Mediation

6.5.1.1 Overview

Mediation is a causal chain of events, where one variable (a mediator variable) at least partially explains (or accounts for) the association between two other variables (the predictor variable and the outcome variable). In mediation, a predictor (`X`) leads to a mediator (`M`), which leads to an outcome (`Y`). Mediation answers the question of, "**Why (or how)** does `X` influence `Y`? A mediator (`M`) is a variable that helps explain the assocation between two other variables, and it answers the question of why/how `X` influences `Y`. That is, the mediator is the variable that helps explain how/why `X` is related to `Y`. In other words, you can think of the mediator as the mechanism that helps explain why `X` has an impact on `Y`. The association between `X` and `Y` gets smaller when accounting for `M`. Visually this can be written as in Figure 6.6:

where `X` is causing `M`, which in turn is causing `Y`. In other words, `X` leads to `M`, and `M` leads to `Y`.

For instance, if we determine that consuming sports drinks improves player performance, we may want to know how/why. That is, what is the mechanism that leads consumption of sports drinks to improve player performance? We might hypothesize that consumption of sports drink helps increase a player's

Figure 6.5 Mediation Versus Moderation Mnemonic.

Figure 6.6 Mediation.

hydration, which in turn will improve the player's performance. In this case, increased hydration mediates (i.e., helps explain or account for) the effect of the sports drink consumption on improved player performance.

Question: Why/how does sports drink consumpion lead players to perform better?
Answer: increased hydration

As a picture, we can draw this assocation as in Figure 6.7:

Figure 6.7 Mediation Example.

6.5.1.2 Types of Mediation

6.5.1.2.1 Full Mediation

When one mechanism fully accounts for the effect of the predictor variable on the outcome variable, this is known as **full mediation**, as depicted in Figure 10.15:

Figure 6.8 Full Mediation.

6.5.1.2.2 Partial Mediation

When a single process partially—but does not fully—accounts for the effect of the predictor variable on the outcome variable; this is known as **partial mediation** and is depicted in Figure 10.16:

6.5.1.2.3 Multiple Mediators

In addition, there can be multiple mediators/mechanisms that account for the effect of a predictor variable on an outcome variable, as depicted in Figure 6.10:

Figure 6.9 Partial Mediation.

Figure 6.10 Multiple Mediators.

6.5.2 Moderation (i.e., Interaction)

6.5.2.1 Overview

Moderation (sometimes called an "interaction"), on the other hand, occurs when there is a variable or condition (`M`; called a "moderator") that changes the assocation between `X` and `Y`. That is, the effect of the predictor variable on the outcome variable differs at different levels of the moderator variable. In these cases, `X` and `M` work together to have an effect on `Y`; here `X` does not have a direct effect on `M`. Moderation answers the question of, "**For whom** does `X` influence `Y`?" If `X` influences `Y` more strongly for some people or in some circumstances, we would say that there is an interaction such that the effect of `X` on `Y` depends on `M`, as depicted in Figure 6.11:

Figure 6.11 Moderation.

For example, if the effect of consuming sports drinks on player performance differs for Quarterbacks and Wide Receivers, the interaction could be depicted in Figures 6.12 and 6.13:

Figure 6.12 Moderation Example: Path Diagram.

An interaction can be identified visually by non-parallel lines at different levels of the moderator. In this example, the player's position moderates the effect consuming sports drinks on player performance. In particular, there is a strong positive association between consuming sports drinks and player performance for Wide Receivers (as evidenced by the upward slope of the best-fit regression line), whereas there is no association between consuming sports drinks and player performance for Quarterbacks (as evidenced by the flat line).

Figure 6.13 Moderation Example: Interaction Graph.

6.6 Levels of Measurement

It is important to know the levels of measurement of your data, because the level(s) of measurement of your data constrain the types of comparisons and analyses that you can meaningfully perform. There are four levels of measurement that any variable can have:

- nominal
- ordinal
- interval
- ratio

Each is described below:

6.6.1 Nominal

A variable is considered nominal if it is composed of qualitative classifications. You cannot meaningfully evaluate whether one number in the variable is larger than another number in the variable because higher numbers do not reflect higher levels of the concept. Examples of nominal variables include:

- sex (e.g., 1 = male; 2 = female)
- race (e.g., 1 = American Indian; 2 = Asian; 3 = Black; 4 = Pacific Islander; 5 = White)
- ethnicity (e.g., 0 = Non-Hispanic/Latino; 1 = Hispanic/Latino)
- zip code
- jersey number

A football player's jersey number is an example of a nominal variable. A jersey number of 7 is not higher on whatever concept of interest compared to a jersey number of 6.

To examine the central tendency of a nominal variable, you can determine the mode, but you cannot calculate a mean or median.

6.6.2 Ordinal

A variable is considered ordinal if the classifications are ordered. However, ordinal variables do not have equally spaced intervals. Examples of ordinal intervals include:

- likert response scales (e.g., 1 = strongly disagree; 2 = disagree; 3 = neutral; 4 = agree; 5 = strongly agree)
- educational attainment (e.g., 1 = no formal education; 2 = elementary school; 3 = middle school; 4 = high school; 5 = college; 6 = graduate degree)
- academic grades on A–F scale (e.g., 1 = A; 2 = B; 3 = C; 4 = D; 5 = F)
- player rank (1 = 1st; 2 = 2nd; 3 = 3rd, etc.)

A football player's fantasy rank is an example of an ordinal variable. A player with a fantasy rank of 1 has a higher rank than a player with a rank of 2, but it is not known how far apart each player is—i.e., the intervals do not all reflect the same distance. For instance, the distance between the top-ranked player and the 2nd-best player might be 30 points, whereas the distance between the 2nd-best player and the 3rd-best player might be 2 points.

To examine the central tendency of ordinal data, the median and mode are most appropriate; however, the mean may be used (unlike for nominal data).

6.6.3 Interval

A variable is considered interval if the classifications are ordered (similar to ordinal data) and have equally spaced intervals (unlike ordinal data). However, interval variables do not have a meaningful zero that reflects absence. Examples of interval data include:

- temperature on the Fahrenheit or Celsius scale
- time of day

For instance, the temperature difference between 80 and 90 degrees Fahrenheit is the same as the temperature difference between 90 and 100 degrees Fahrenheit. However, 0 degrees Fahrenheit does not reflect absence of temperature/heat.

Interval data can be meaningfully added or subtracted. For instance, if a game starts at 4 pm and ends at 7 pm, you know the game lasted 3 hours ($7 - 4 = 3$). However, interval data cannot be meaningfully multiplied or divided. For instance, 100 degrees Fahrenheit is not twice as hot as 50 degrees Fahrenheit.

To examine the central tendency of interval data, you can compute the mean, median, or mode.

6.6.4 Ratio

A variable is considered ratio if the classifications are ordered (similar to ordinal data), have equally spaced intervals (like interval data), and have an absolute zero point that reflects absence of the concept. Examples of ratio data include:

- temperature on the Kelvin scale
- height
- weight
- age
- distance
- speed
- volume
- time elapsed
- income
- stock price
- years of formal education
- points in football

For instance, points in football has order, equally spaced intervals, and an absolute zero—a team cannot score less than zero points, and zero points reflects absence of points (though it could be argued to be interval data because zero points does not reflect absence of skill.)

Ratio data can be meaningfully added, subtracted, multiplied, or divided. A player who weighs 350 pounds weighs twice as much as someone who weighs 175 pounds.

To examine the central tendency of ratio data, you can compute the mean, median, or mode.

6.7 Psychometrics

Below, I provide brief discussions of various aspects of measurement reliability and validity. For more information on these and other aspects of psychometrics, see Petersen (2024b) and Petersen (2024c).

6.7.1 Measurement Reliability

The *reliability* of a measure's scores deals with the *consistency* of measurement. This book focuses on the following types of reliability:

- test–retest reliability
- inter-rater reliability
- intra-rater reliability
- internal consistency
- parallel-forms reliability

For more information on these and other aspects of reliability, see https://isaactpetersen.github.io/Principles-Psychological-Assessment/reliability.html (Petersen, 2024b, 2024c).

6.7.1.1 Test–Retest Reliability

Test–retest reliability evaluates the consistency of scores across time. For a construct that is expected to be stable across time (e.g., hand size in adults), we would expect our measurements to be consistent across time. The consistency of scores across time can be examined in terms of relative or absolute test–retest reliability. Relative test–retest reliability—i.e., the consistency of

individual differences across time—is commonly evaluated using the coefficient of stability (i.e., the Pearson correlation coefficent). Absolute test–retest reliability—i.e., the absolute consistency of people's scores across time—is commonly evaluated using the coefficient of repeatability.

6.7.1.2 Inter-Rater Reliability

Inter-rater reliability evaluates the consistency of scores across raters. For instance, if we have a strong measure for assessing college players' aptitude to succeed in the NFL, the measure should yield a similar score for a given player regardless of which (trained) rater (e.g., coach or talent scout) uses it to rate the player. The consistency of scores across raters is commonly evaluated using the intraclass correlation coefficient (for continuous variables) and Cohen's kappa (κ; for categorical variables).

6.7.1.3 Intra-Rater Reliability

Intra-rater reliability evaluates the consistency of scores within a given rater. If we have a strong measure for assessing college players' aptitude to succeed in the NFL, the measure should yield a similar score for a given player from the same (trained) rater (e.g., coach or talent scout) each time they rate the same player (assuming the player's aptitude has not changed). The consistency of scores within raters can be evaluated using similar approaches as those evaluating inter-rater reliability.

6.7.1.4 Internal Consistency

Internal consistency evaluates the consistency of scores across items within a measure. If we develop a strong questionnaire measure to assess a college players' aptitude to succeed in the NFL, the scores should be relatively consistent across items. The consistency of scores across items within a measure is commonly evaluated using Cronbach's alpha (α) or McDonald's omega (ω).

6.7.1.5 Parallel-Forms Reliability

Parallel-forms reliability evaluates the consistency of scores across different but equivalent forms of a measure. If we develop two equivalent versions of the Wonderlic Contemporary Cognitive Ability Test (Form A and Form B) so that players sitting next to each other do not receive the same items, we would expect a player's score on Form A would be similar to their score on Form B. Parallel-forms reliability is is commonly evaluated using the coefficient of equivalence (i.e., the Pearson correlation coefficent).

6.7.2 Measurement Validity

The *validity* of a measure's scores deals with the *accuracy* of measurement. This book focuses on the following types of validity:

- face validity
- content validity
- criterion-related validity
  - concurrent (criterion-related) validity
  - predictive (criterion-related) validity
- construct validity
- convergent validity
- discriminant validity
- incremental validity
- ecological validity

For more information on these and other aspects of validity, see https://isaactpetersen.github.io/Principles-Psychological-Assessment/validity.html (Petersen, 2024b, 2024c).

6.7.2.1 Face Validity

Face validity evaluates the extent to which a measure "looks like" (on its face) it assesses the construct of interest. For instance, if a measure is developed to assess aptitude of Wide Receivers for the position, it would be considered to have face validity if everyday (lay) people believe that it assesses aptitude for being a successful Wide Receiver.

6.7.2.2 Content Validity

Content validity evaluates the extent to which the measure assesses the full breadth of the content, as determined by context experts. For the measure to have content validity, it should not have gaps (missing content facets) or intrusions (facets of other constructs). For instance, a strong measure for assessing a player's aptitude to succeed in the NFL might need to include a player's speed, strength, size, lateral quickness, etc. If the measure is missing their speed, this would be a content gap. If the measure assesses a construct-irrelevant facet (e.g., their attractiveness), this would be a content intrusion.

6.7.2.3 Criterion-Related Validity

Criterion-related validity evaluates the extent to which the measure's scores are related to meaningful variables of interest. Criterion-related validity is commonly evaluated using a Pearson correlation or some form of regression.

There are two types of criterion-related validity:

- concurrent (criterion-related) validity
- predictive (criterion-related) validity

6.7.2.3.1 Concurrent (Criterion-Related) Validity

Concurrent criterion-related validity (aka concurrent validity) evaluates the extent to which the measure's scores are related to meaningful variables of interest assessed at the same point in time. That is, concurrent validity could evaluate whether current player statistics (e.g., passing yards) are associated with their fantasy points.

6.7.2.3.2 Predictive (Criterion-Related) Validity

Predictive criterion-related validity (aks predictive validity) evaluates the extent to which the measure's scores are related to meaningful variables of interest that are assessed at a later point in time. For example, predictive validity could evaluate whether scores on the measure we developed to assess a player's aptitude to succeed in the NFL predicts later performance in the NFL.

6.7.2.4 Construct Validity

Construct validity evaluates the extent to which the measure's scores accurately assess the construct of interest. If we develop a measure with intent to assess aptitude for being a successful Running Back, and it appears to more accurately assess aptitude for being a successful Wide Receiver, then our measure has poor construct validity for assessing aptitude for being a successful Running Back. Construct validity subsumes convergent and discriminant validity, in addition to all of the other forms of measurement validity.

6.7.2.5 Convergent Validity

Convergent validity evaluates the extent to which the measure's scores are related to other measures of the same construct. For instance, if we develop a new measure to assess intelligence, its scores should be related to scores from other measures designed to assess intelligence (e.g., Wonderlic Contemporary Cognitive Ability Test).

6.7.2.6 Discriminant Validity

Discriminant validity evaluates the extent to which the measure's scores are unrelated to measures of the different constructs. For instance, if we develop a new measure to assess intelligence, its scores should be less strongly associated with measures of other constructs (e.g., measures of happiness).

6.7.2.7 Incremental Validity

Incremental validity evalutes the extent to which the measure's scores provide an increase in predictive accuracy compared to other information that is easily and cheaply available. That is, in order to be useful, a strong measure should tell us something that we did not already know. For instance, if we develop a strong measure of intelligence, it should result in increased predictive accuracy (for success in the NFL) compared to when just relying on the Wonderlic Contemporary Cognitive Ability Test.

6.7.2.8 Ecological Validity

Ecological validity evaluates the extent to which the measures' scores are indicative of the behavior of a person in the natural environment. For instance, measures of a players' speed during a game has higher ecological validity (and is more predictive of their performance) than their speed during the NFL Combine (Lyons et al., 2011). For instance, compared to tests of speed, power, and agility at the NFL Combine, collegiate performance is a stronger predictor of performance in the NFL (Lyons et al., 2011). That is, previous sports performance is the best predictor of future performance (for a review, see Den Hartigh et al., 2018).

6.7.3 Reliability vs Validity

Reliability and validity are different but related. Reliability refers to the *consistency* of scores, whereas accuracy refers to the *accuracy* of scores. Validity depends on reliability. Reliability is necessary—but insufficient for—validity. That is, consistency is necessary—but insufficient for—accuracy. As depicted in Figure 6.14, a measure can be no more valid than it is reliable. A measure can be consistent but inaccurate; however, a measure cannot be accurate but inconsistent.

Figure 6.14 Reliability Versus Validity.

6.8 Conclusion

There are various types of research designs. Each type of research design differs in the extent to which it supports the ability to draw causal inferences (internal validity) versus the extent to which it supports the ability to identify processes that generalize to the real-world (external validity). In addition, it is important to understand the distinction between sample and population, and the distinction between mediation and moderation. It is also important to consider the levels of measurement used because they constrain the types of analyses that may be performed. In addition, it is important to consider the psychometrics of measurements, including multiple aspects of reliability (consistency) and validity (accuracy).

7

Basic Statistics

7.1 Getting Started

7.1.1 Load Packages

```
library("petersenlab")
library("DescTools")
library("pwr")
library("pwrss")
library("WebPower")
library("grid")
library("tidyverse")
```

7.2 Descriptive Statistics

Descriptive statistics are used to describe data. For instance, they may be used to describe the center, spread, or shape of the data. There are various indices of each.

7.2.1 Center

Indices to describe the *center* (central tendency) of a variable's data include:

- mean
- median
- Hodges-Lehmann statistic (aka pseudomedian)
- mode
- weighted mean

- weighted median

The mean of X (written as: $\bar{X}$) is calculated as in Equation 7.5:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + ... + X_n}{n} \tag{7.1}$$

```
exampleValues <- c(0, 0, 10, 15, 20, 30, 1000)
exampleValues_mean <- apa(mean(exampleValues), 2)
```

That is, to compute the mean, sum all of the values and divide by the number of values (n). One issue with the mean is that it is sensitive to extreme (outlying) values. For instance, the mean of the values of 0, 0, 10, 15, 20, 30, and 1000 is 153.57.

```
exampleValues_median <- median(exampleValues)
```

The median is determined as the value at the 50th percentile (i.e., the value that is higher than 50% of the values and is lower than the other 50% of values). Compared to the mean, the median is less influenced by outliers. The median of the values of 0, 0, 10, 15, 20, 30, and 1000 is 15.

```
exampleValues_pseudomedian <- DescTools::HodgesLehmann(exampleValues)
```

The Hodges-Lehmann statistic (aka pseudomedian) is computed as the median of all pairwise means, and it is also robust to outliers. The pseudomedian of the values of 0, 0, 10, 15, 20, 30, and 1000 is 15.

```
exampleValues_mode <- petersenlab::Mode(exampleValues)
```

The mode is the most common/frequent value. The mode of the values of 0, 0, 10, 15, 20, 30, and 1000 is 0. The `petersenlab`[1] package (Petersen, 2024a) contains the `Mode()` function for computing the mode of a set of data.

If you want to give some values more weight to others, you can calculate a weighted mean and a weighted median (or other quantile), while assigning a weight to each value.

Below is R code to estimate each:

[1] https://github.com/DevPsyLab/petersenlab

```
#mean(data, na.rm = TRUE)
#median(data, na.rm = TRUE)
#DescTools::HodgesLehmann(exampleValues, na.rm = TRUE)
#petersenlab::Mode(exampleValues)
#weighted.mean(data, weights, na.rm = TRUE)
#DescTools::Quantile(data, weights, na.rm = TRUE)
```

7.2.2 Spread

Indices to describe the *spread* (variability) of a variable's data include:

- standard deviation
- variance
- range
- minimum and maximum
- interquartile range (IQR)
- median absolute deviation

The (sample) variance of X (written as: s^2) is calculated as in Equation 7.2:

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1} \tag{7.2}$$

where X_i is each data point, $\bar{X}$ is the mean of X, and n is the number of data points.

The (sample) standard deviation of X (written as: s) is calculated as in Equation 7.3:

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} \tag{7.3}$$

The range is calculated of X is calculated as in Equation 7.4:

$$\text{range} = \text{maximum} - \text{minimum} \tag{7.4}$$

The interquartile range (IQR) is calculated as in Equation 7.5:

$$\text{IQR} = Q_3 - Q_1 \tag{7.5}$$

where Q_3 is the score at the third quartile (i.e., 75th percentile), and Q_1 is the score at the first quartile (i.e., 25th percentile).

The median absolute deviation (MAD) is the median of all deviations from the median, and is calculated as in Equation 7.6:

$$\text{MAD} = \text{median}(|X_i - \tilde{X}|) \tag{7.6}$$

where $\tilde{X}$ is the median of `x`. Compared to the standard deviation, the median absolute deviation is more robust to outliers.

Below is R code to estimate each:

7.2.3 Shape

Indices to describe the *shape* of a variable's data include:

- skewness
- kurtosis

Below is R code to estimate each:

7.2.4 Combination

To estimate multiple indices of center, spread, and shape of the data, you can use the following code:

```
#psych::describe(mydata)

#mydata %>%
#  summarise(across(
#      everything(),
#      .fns = list(
#        n = ~ length(na.omit(.)),
#        missingness = ~ mean(is.na(.)) * 100,
#        M = ~ mean(., na.rm = TRUE),
#        SD = ~ sd(., na.rm = TRUE),
#        min = ~ min(., na.rm = TRUE),
#        max = ~ max(., na.rm = TRUE),
#        range = ~ max(., na.rm = TRUE) - min(., na.rm = TRUE),
#        IQR = ~ IQR(., na.rm = TRUE),
#        MAD = ~ mad(., na.rm = TRUE),
#        median = ~ median(., na.rm = TRUE),
#        pseudomedian = ~ DescTools::HodgesLehmann(., na.rm = TRUE),
#        mode = ~ petersenlab::Mode(., multipleModes = "mean"),
#        skewness = ~ psych::skew(., na.rm = TRUE),
```

```
#        kurtosis = ~ psych::kurtosi(., na.rm = TRUE)),
#      .names = "{.col}.{.fn}")) %>%
#    pivot_longer(
#      cols = everything(),
#      names_to = c("variable","index"),
#      names_sep = "\\.") %>%
#    pivot_wider(
#      names_from = index,
#      values_from = value)
```

7.3 Scores and Scales

There are many different types of scores and scales. This book focuses on raw scores and *z*-scores. For information on other scores and scales, including percentile ranks, *T*-scores, standard scores, scaled scores, and stanine scores, see here: https://isaactpetersen.github.io/Principles-Psychological-Assessment/scoresScales.html#scoreTransformation (Petersen, 2024c).

7.3.1 Raw Scores

Raw scores are the original data on the original metric. Thus, raw scores are considered *unstandardized*. For example, raw scores that represent the players' age may range from 20 to 40. Raw scores depend on the construct and unit; thus raw scores may not be comparable across variables.

7.3.2 *z* Scores

z scores have a mean of zero and a standard deviation of one. z scores are frequently used to render scores across variables more comparable. Thus, z scores are considered a form of a *standardized* score.

z scores are calculated using Equation 7.7:

$$z = \frac{X - \bar{X}}{\sigma} \tag{7.7}$$

where X is the observed score, $\bar{X}$ is the mean observed score, and σ is the standard deviation of the observed scores.

You can easily convert a variable to a z score using the `scale()` function:

```
scale(variable)
```

With a standard normal curve, 68% of scores fall within one standard deviation of the mean. 95% of scores fall within two standard deviations of the mean. 99.7% of scores fall within three standard deviations of the mean.

The area under a normal curve within one standard deviation of the mean is calculated below using the `pnorm()` function, which calculates the cumulative density function for a normal curve.

```
stdDeviations <- 1

pnorm(stdDeviations) - pnorm(stdDeviations * -1)
```

```
[1] 0.6826895
```

The area under a normal curve within one standard deviation of the mean is depicted in Figure 7.1.

```
x <- seq(-4, 4, length = 200)
y <- dnorm(x, mean = 0, sd = 1)
plot(x, y, type = "l",
     xlab = "z Score",
     ylab = "Normal Density")

x <- seq(stdDeviations * -1, stdDeviations, length = 100)
y <- dnorm(x, mean = 0, sd = 1)
polygon(c(stdDeviations * -1, x, stdDeviations),
        c(0, y, 0),
        col = "blue")
```

Figure 7.1 Density of Standard Normal Distribution. The blue region represents the area within one standard deviation of the mean.

The area under a normal curve within two standard deviations of the mean is calculated below:

```
stdDeviations <- 2

pnorm(stdDeviations) - pnorm(stdDeviations * -1)
```

```
[1] 0.9544997
```

The area under a normal curve within two standard deviations of the mean is depicted in Figure 7.2.

```
x <- seq(-4, 4, length = 200)
y <- dnorm(x, mean = 0, sd = 1)
plot(x, y, type = "l",
     xlab = "z Score",
     ylab = "Normal Density")

x <- seq(stdDeviations * -1, stdDeviations, length = 100)
y <- dnorm(x, mean = 0, sd = 1)
polygon(c(stdDeviations * -1, x, stdDeviations),
        c(0, y, 0),
        col = "blue")
```

Figure 7.2 Density of Standard Normal Distribution. The blue region represents the area within two standard deviations of the mean.

The area under a normal curve within three standard deviations of the mean is calculated below:

```
stdDeviations <- 3

pnorm(stdDeviations) - pnorm(stdDeviations * -1)
```

```
[1] 0.9973002
```

The area under a normal curve within three standard deviations of the mean is depicted in Figure 7.3.

```
x <- seq(-4, 4, length = 200)
y <- dnorm(x, mean = 0, sd = 1)
plot(x, y, type = "l",
     xlab = "z Score",
     ylab = "Normal Density")

x <- seq(stdDeviations * -1, stdDeviations, length = 100)
y <- dnorm(x, mean = 0, sd = 1)
polygon(c(stdDeviations * -1, x, stdDeviations),
        c(0, y, 0),
        col = "blue")
```

Figure 7.3 Density of Standard Normal Distribution. The blue region represents the area within three standard deviations of the mean.

If you want to determine the z score associated with a particular percentile in a normal distribution, you can use the `qnorm()` function. For instance, the z score associated with the 37th percentile is:

```
qnorm(.37)
```

```
[1] -0.3318533
```

7.4 Inferential Statistics

Inferential statistics are used to draw inferences regarding whether there is (a) a difference in level on variable across groups or (b) an association between variables. For instance, inferential statistics may be used to evaluate whether Quarterbacks tend to have longer careers compared to Running Backs. Or, they could be used to evaluate whether number of carries is associated with injury likelihood. To apply inferential statistics, we make use of the null hypothesis (H_0) and the alternative hypothesis (H_1).

7.4.1 Null Hypothesis Significance Testing

To draw statistical inferences, the frequentist statistics paradigm leverages null hypothesis significance testing. Frequentist statistics is the most widely used statistical paradigm. However, frequentist statistics is not the only statistical paradigm. Other statistical paradigms exist, including Bayesian statistics, which is based on Bayes' theorem. This chapter focuses on the frequentist approach to hypothesis testing, known as null hypothesis significance testing. We discuss Bayesian statistics in Chapter 13.

7.4.1.1 Null Hypothesis (H_0)

When testing whether there are differences in level across groups on a variable of interest, the null hypothesis (H_0) is that there is no difference in level across groups. For instance, when testing whether Quarterbacks tend to have longer careers compared to Running Backs, the null hypothesis (H_0) is that Quarterbacks do not systematically differ from Running Backs in the length of their career.

When testing whether there is an association between variables, the null hypothesis (H_0) is that there is no association between the variables. For instance, when testing whether number of carries is associated with injury likelihood, the null hypothesis (H_0) is that there is no association between number of carries and injury likelihood.

7.4.1.2 Alternative Hypothesis (H_1)

The alternative hypothesis (H_1) is the researcher's hypothesis that they want to evaluate. An alternative hypothesis (H_1) might be directional (i.e., one-sided) or non-directional (i.e., two-sided).

Directional hypotheses specify a particular direction, such as which group will have larger scores or which direction (positive or negative) two variables will be associated. Examples of directional hypotheses include:

- Quarterbacks have longer careers compared to Running Backs
- Number of carries is positively associated with injury likelihood

Non-directional hypotheses do not specify a particular direction. For instance, non-directional hypotheses may state that two groups differ but do not specify which group will have larger scores. Or, non-directional hypotheses may state that two variables are associated but do not state what the sign is of the association—i.e., positive or negative. Examples of non-directional hypotheses include:

- Quarterbacks differ in the length of their careers compared to Running Backs
- Number of carries is associated with injury likelihood

7.4.1.3 Statistical Significance

In science, statistical significance is evaluated with the p-value. The p-value does not represent the probability that you observed the result by chance. The p-value represents a conditional probability—it examines the probability of one event given another event. In particular, the p-value evaluates the likelihood that you would detect a result as at least as extreme as the one observed (in terms of the magnitude of the difference or of the association) given that the null hypothesis (H_0) is true.

This can be expressed in conditional probability notation, $P(A|B)$, which is the probability (likelihood) of event A occurring given that event B occurred (or given condition B).

The conditional probability notation for a left-tailed directional test (i.e., Quarterbacks have shorter careers than Running Backs; or number of carries is negatively associated with injury likelihood) is in Equation 7.8.

$$p\text{-value} = P(T \leq t|H_0) \tag{7.8}$$

where T is the test statistic of interest (e.g., the distribution of t-, $r-$, or F values, depending on the test) and t is the observed test statistic (e.g., t-, $r-$, or F-coefficient, depending on the test).

The conditional probability notation for a right-tailed directional test (i.e., Quarterbacks have longer careers than Running Backs; or number of carries is positively associated with injury likelihood) is in Equation 7.9.

$$p\text{-value} = P(T \geq t|H_0) \tag{7.9}$$

The conditional probability notation for a two-tailed non-directional test (i.e., Quarterbacks differ in the length of their careers compared to Running Backs; or number of carries is associated with injury likelihood) is in Equation 7.10.

$$p\text{-value} = 2 \times \min(P(T \leq t|H_0), P(T \geq t|H_0)) \tag{7.10}$$

where `min(a, b)` is the smaller number of `a` and `b`.

If the distribution of the test statistic is symmetric around zero, the p-value for the two-tailed non-directional test simplifies to Equation 7.11.

$$p\text{-value} = 2 \times P(T \geq |t||H_0) \tag{7.11}$$

Nevertheless, to be conservative (i.e., to avoid false positive/Type I errors), many researchers use two-tailed p-values regardless whether their hypothesis is one- or two-tailed.

For a test of group differences, the p-value evaluates the likelihood that you would observe a difference as large or larger than the one you observed between the groups if there were no systematic difference between the groups, as depicted in Figure 7.4. For instance, when evaluating whether Quarterbacks have longer careers than Running Backs, and you observed a mean difference of 0.03 years, the p-value evaluates the likelihood that you would observe a difference as larger or larger than 0.03 years between the groups if Quarterbacks do not differ from Running Backs in terms of the length of their career.

```
set.seed(52242)

nObserved <- 1000
nPopulation <- 1000000

observedGroups <- data.frame(
  score = c(rnorm(nObserved, mean = 47, sd = 3), rnorm(nObserved, mean = 52, sd = 3)),
  group = as.factor(c(rep("Group 1", nObserved), rep("Group 2", nObserved)))
)

populationGroups <- data.frame(
  score = c(rnorm(nPopulation, mean = 50, sd = 3.03), rnorm(nPopulation, mean = 50, sd = 3)),
  group = as.factor(c(rep("Group 1", nPopulation), rep("Group 2", nPopulation)))
)

ggplot2::ggplot(
  data = observedGroups,
  mapping = aes(
    x = score,
    fill = group,
    color = group
  )
) +
  geom_density(alpha = 0.5) +
  scale_color_manual(values = c("red", "blue")) +
  scale_fill_manual(values = c("red","blue")) +
  geom_vline(xintercept = mean(observedGroups$score[which(observedGroups$group == "Group 1")])) +
  geom_vline(xintercept = mean(observedGroups$score[which(observedGroups$group == "Group 2")])) +
  ggplot2::labs(
    x = "Score",
    y = "Frequency",
    title = "What is the probability my data would look like this..."
  ) +
```

```
  ggplot2::theme_classic(
    base_size = 16) +
  ggplot2::theme(
    legend.title = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank(),
    #plot.title.position = "plot"
    legend.position = "inside",
    legend.margin = margin(0, 0, 0, 0),
    legend.justification.top = "left",
    legend.justification.left = "top",
    legend.justification.bottom = "right",
    legend.justification.inside = c(1, 1),
    legend.location = "plot")

ggplot2::ggplot(
  data = populationGroups,
  mapping = aes(
    x = score,
    fill = group,
    color = group
  )
) +
  geom_density(alpha = 0.5) +
  scale_color_manual(values = c("red", "blue")) +
  scale_fill_manual(values = c("red","blue")) +
  geom_vline(xintercept = mean(populationGroups$score[which(populationGroups$group == "Group 1")])
  geom_vline(xintercept = mean(populationGroups$score[which(populationGroups$group == "Group 2")])
  ggplot2::labs(
    x = "Score",
    y = "Frequency",
    title = "...if in the population, the groups were really this:"
  ) +
  ggplot2::theme_classic(
    base_size = 16) +
  ggplot2::theme(
    legend.title = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank(),
    #plot.title.position = "plot",
```

```
    legend.position = "inside",
    legend.margin = margin(0, 0, 0, 0),
    legend.justification.top = "left",
    legend.justification.left = "top",
    legend.justification.bottom = "right",
    legend.justification.inside = c(1, 1),
    legend.location = "plot")
```

(a) What is the probability my data would look like this… **(b)** …if in the population, the groups were really this?

Figure 7.4 Interpretation of p-Values When Examining The Differences Between Groups. The vertical black lines reflect the group means.

For a test of whether two variables are associated, the p-value evaluates the likelihood that you would observe an association as strong or stronger than the one you observed between the groups if there were no association between the variables, as depicted in Figure 7.5. For instance, when evaluating whether number of carries is positively associated with injury likelihood, and you observed a correlation coefficient of $r = .25$ between number of carries and injury likelihood, the p-value evaluates the likelihood that you would observe a correlation as strong or stronger than $r = .25$ between the variables if number of carries is not associated with injury likelihood.

```
set.seed(52242)

observedCorrelation <- 0.9

correlations <- data.frame(criterion = rnorm(2000))
correlations$sample <- NA
correlations$sample[1:100] <- complement(correlations$criterion[1:100], observedCorrelation)
correlations$population <- complement(correlations$criterion, 0)

ggplot2::ggplot(
  data = correlations,
```

```
  mapping = aes(
    x = sample,
    y = criterion
  )
) +
  geom_point() +
  geom_smooth(method = "lm") +
  scale_x_continuous(
    limits = c(-3.5,3)
  ) +
  annotate(
    x = 0,
    y = 4,
    label = paste("italic(r) != ", 0, sep = ""),
    parse = TRUE,
    geom = "text",
    size = 7) +
  ggplot2::labs(
    x = "Predictor Variable",
    y = "Outcome Variable",
    title = "What is the probability my data would look like this..."
  ) +
  ggplot2::theme_classic(
    base_size = 16) +
  ggplot2::theme(
    legend.title = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())

ggplot2::ggplot(
  data = correlations,
  mapping = aes(
    x = population,
    y = criterion
  )
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE) +
  scale_x_continuous(
    limits = c(-2.5,2.5)
```

```
  ) +
  annotate(
    x = 0,
    y = 4,
    label = paste("italic(r) == '", "0.00", "'", sep = ""),
    parse = TRUE,
    geom = "text",
    size = 7) +
  ggplot2::labs(
    x = "Predictor Variable",
    y = "Outcome Variable",
    title = "...if in the population, the association was really this:"
  ) +
  ggplot2::theme_classic(
    base_size = 16) +
  ggplot2::theme(
    legend.title = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())
```

(a) What is the probability my data would look like this…

(b) …if in the population, the association was really this?

Figure 7.5 Interpretation of p-Values When Examining The Association Between Variables.

Using what is called null-hypothesis significance testing (NHST), we consider an effect to be *statistically significant* if the p-value is less than some threshold, called the *alpha level*. In science, we typically want to be conservative because a false positive (i.e., Type I error) is considered more problematic than a false negative (i.e., Type II error). That is, we would rather say an effect does not exist when it really does than to say an effect does exist when it really does not. Thus, we typically set the alpha level to a low value, commonly .05.

Then, we would consider an effect to be *statistically significant* if the p-value is less than .05. That is, there is a small chance (5%; or 1 in 20 times) that we would observe an effect at least as extreme as the effect observed, if the null hypothesis were true. So, you might expect around 5% of tests where the null hypothesis is true to be statistically significant just by chance. We could lower the rate of Type II (i.e., false negative) errors—i.e., we could detect more effects—if we set the alpha level to a higher value (e.g., .10); however, raising the alpha level would raise the possibility of Type I (false positive) errors.

If the p-value is less than .05, we reject the null hypothesis (H_0) that there was no difference or association. Thus, we conclude that there was a statistically significant (non-zero) difference or association. If the p-value is greater than .05, we fail to reject the null hypothesis; the difference/association was not statistically significant. Thus, we do not have confidence that there was a difference or association. However, we do not accept the null hypothesis; it could be there we did not observe an effect because we did not have adequate power to detect the effect—e.g., if the effect size was small, the data were noisy, and the sample size was small and/or unrepresentative.

There are four general possibilities of decision making outcomes when performing null-hypothesis significance testing:

1. We (correctly) reject the null hypothesis when it is in fact false ($1-\beta$). This is a true positive. For instance, we may correctly determine that Quarterbacks have longer careers than Running Backs.
2. We (correctly) fail to reject the null hypothesis when it is in fact true ($1-\alpha$). This is a true negative. For instance, we may correctly determine that Quarterbacks do not have longer careers than Running Backs.
3. We (incorrectly) reject the null hypothesis when it is in fact true (α). This is a false positive. When performing null hypothesis testing, a false positive is known as a Type I error. For instance, we may incorrectly determine that Quarterbacks have longer careers than Running Backs when, in fact, Quarterbacks and Running Backs do not differ in their career length.
4. We (incorrectly) fail to reject the null hypothesis when it is in fact false (β). This is a false negative. When performing null hypothesis testing, a false negative is known as a Type II error. For instance, we may incorrectly determine that Quarterbacks and Running Backs do not differ in their career length when, in fact, Quarterbacks have longer careers than Running Backs.

A two-by-two confusion matrix for null-hypothesis significance testing is in Figure 7.6.

		Decision	
		Reject H_0	Fail to reject H_0
Truth	H_0 false	Correct True Positive 1 – β (“power”)	Type II error False Negative beta (β)
	H_0 true	Type I error False Positive alpha (α)	Correct True Negative 1 – α

Figure 7.6 A Two-by-Two Confusion Matrix for Null-Hypothesis Significance Testing.

In statistics, *power* is the probability of detecting an effect, if, in fact, the effect exists. Otherwise said, power is the probability of rejecting the null hypothesis, if, in fact, the null hypothesis is false. Power is influenced by several variables:

- the sample size (N): the larger the N, the greater the power
  - for group comparisons, the power depends on the sample size of each group
- the effect size: the larger the effect, the greater the power
  - for group comparisons, larger effect sizes reflect:
    * larger between-group variance, and
    * smaller within-group variance (i.e., strong measurement precision, i.e., reliability)
- the alpha level: the researcher specifies the alpha level (though it is typically set at .05); the higher the alpha level, the greater the power; however, the higher we set the alpha level, the higher the likelihood of Type I errors (false positives)
- one- versus two-tailed tests: one-tailed tests have higher power than two-tailed tests
- within-subject versus between-subject comparisons: within-subject designs tend to have greater power than between-subject designs

A plot of statistical power is in Figure 7.7.

```
m1 <- 0  # mu H0
sd1 <- 1.5 # sigma H0
m2 <- 3.5 # mu HA
```

```
sd2 <- 1.5 # sigma HA

z_crit <- qnorm(1-(0.05/2), m1, sd1)

# set length of tails
min1 <- m1-sd1*4
max1 <- m1+sd1*4
min2 <- m2-sd2*4
max2 <- m2+sd2*4
# create x sequence
x <- seq(min(min1,min2), max(max1, max2), .01)
# generate normal dist #1
y1 <- dnorm(x, m1, sd1)
# put in data frame
df1 <- data.frame("x" = x, "y" = y1)
# generate normal dist #2
y2 <- dnorm(x, m2, sd2)
# put in data frame
df2 <- data.frame("x" = x, "y" = y2)

# Alpha polygon
y.poly <- pmin(y1,y2)
poly1 <- data.frame(x=x, y=y.poly)
poly1 <- poly1[poly1$x >= z_crit, ]
poly1<-rbind(poly1, c(z_crit, 0))  # add lower-left corner

# Beta polygon
poly2 <- df2
poly2 <- poly2[poly2$x <= z_crit,]
poly2<-rbind(poly2, c(z_crit, 0))  # add lower-left corner

# power polygon; 1-beta
poly3 <- df2
poly3 <- poly3[poly3$x >= z_crit,]
poly3 <-rbind(poly3, c(z_crit, 0))  # add lower-left corner

# combine polygons.
poly1$id <- 3 # alpha, give it the highest number to make it the top layer
poly2$id <- 2 # beta
poly3$id <- 1 # power; 1 - beta
poly <- rbind(poly1, poly2, poly3)
poly$id <- factor(poly$id,  labels=c("power","beta","alpha"))

# plot with ggplot2
```

```
ggplot(poly, aes(x,y, fill=id, group=id)) +
  geom_polygon(show.legend=F, alpha=I(8/10)) +
  # add line for treatment group
  geom_line(data=df1, aes(x,y, color="H0", group=NULL, fill=NULL), linewidth=1.5, show_guide=F) +
  # add line for treatment group. These lines could be combined into one dataframe.
  geom_line(data=df2, aes(color="HA", group=NULL, fill=NULL),linewidth=1.5, show_guide=F) +
  # add vlines for z_crit
  geom_vline(xintercept = z_crit, linewidth=1, linetype="dashed") +
  # change colors
  scale_color_manual("Group",
                     values= c("HA" = "#981e0b","H0" = "black")) +
  scale_fill_manual("test", values= c("alpha" = "#0d6374","beta" = "#be805e","power"="#7cecee")) +
  # beta arrow
  annotate("segment", x=0.1, y=0.045, xend=1.3, yend=0.01, arrow = arrow(length = unit(0.3, "cm"))
  annotate("text", label="beta", x=0, y=0.05, parse=T, size=8) +
  # alpha arrow
  annotate("segment", x=4, y=0.043, xend=3.4, yend=0.01, arrow = arrow(length = unit(0.3, "cm")),
  annotate("text", label="frac(alpha,2)", x=4.2, y=0.05, parse=T, size=8) +
  # power arrow
  annotate("segment", x=6, y=0.2, xend=4.5, yend=0.15, arrow = arrow(length = unit(0.3, "cm")), li
  annotate("text", label=expression(paste(1-beta, "  (\"power\")")), x=6.1, y=0.21, parse=T, size=
  # H_0 title
  annotate("text", label="H[0]", x=m1, y=0.28, parse=T, size=8) +
  # H_a title
  annotate("text", label="H[1]", x=m2, y=0.28, parse=T, size=8) +
  ggtitle("Statistical Power") +
  # remove some elements
  theme(
    panel.grid.minor = element_blank(),
    panel.grid.major = element_blank(),
    panel.background = element_blank(),
    plot.background = element_rect(fill="white"),
    panel.border = element_blank(),
    axis.line = element_blank(),
    axis.text.x = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks = element_blank(),
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    plot.title = element_text(size=22))
```

Figure 7.7 Statistical Power (Adapted from Kristoffer Magnusson: https://rpsychologist.com/creating-a-typical-textbook-illustration-of-statistical-power-using-either-ggplot-or-base-graphics; archived at https://perma.cc/FG3J-85L6). The dashed line represents the critical value or threshold.

Interactive visualizations by Kristoffer Magnusson on p-values and null-hypothesis significance testing are below:

- https://rpsychologist.com/pvalue/ (archived at https://perma.cc/JP9F-9ZVY)
- https://rpsychologist.com/d3/pdist/ (archived at https://perma.cc/BE96-8LSJ)
- https://rpsychologist.com/d3/nhst/ (archived at https://perma.cc/ZU9A-37F3)

Twelve misconceptions about p-values (Goodman, 2008) are in Table 7.1.

Table 7.1 Twelve Misconceptions About p-Values from Goodman (2008). Goodman also provides a discussion about why each statement is false.

Number	Misconception
1	If $p = .05$, the null hypothesis has only a 5% chance of being true.
2	A nonsignificant difference (eg, $p > .05$) means there is no difference between groups.

Number	Misconception
3	A statistically significant finding is clinically important.
4	Studies with p-values on opposite sides of .05 are conflicting.
5	Studies with the same p-value provide the same evidence against the null hypothesis.
6	$p = .05$ means that we have observed data that would occur only 5% of the time under the null hypothesis.
7	$p = .05$ and $p < .05$ mean the same thing.
8	p-values are properly written as inequalities (e.g., "$p \leq .05$" when $p = .015$).
9	$p = .05$ means that if you reject the null hypothesis, the probability of a Type I error is only 5%.
10	With a $p = .05$ threshold for significance, the chance of a Type I error will be 5%.
11	You should use a one-sided p-value when you don't care about a result in one direction, or a difference in that direction is impossible.
12	A scientific conclusion or treatment policy should be based on whether or not the p-value is significant.

That is, the p-value is not:

- the probability that the effect was due to chance
- the probability that the null hypothesis is true
- the size of the effect
- the importance of the effect
- whether the effect is true, real, or causal

Statistical significance involves the *consistency* of an effect/association/difference; it suggests that the association/difference is reliably non-zero. However, just because something is statistically significant does not mean that it is important. For instance, consider that we discover that players who consume sports drink before a game tend to perform better than players who do not ($p < .05$). However, what if consumption of sports drinks is associated with an average improvement of 0.002 points per game. A small effect such as this might be detectable with a large sample size. This effect would be considered to be reliable/consistent because it is statistically significant. However, such an effect is so small that it results in differences that are not practically important. Thus, in addition to statistical significance, it is also important to consider practical significance.

7.4.2 Practical Significance

Practical significance deals with how large or important the effect/association/difference is. It is based on the magnitude of the effect, called the *effect size*. Effect size can be quantified in various ways including:

- Cohen's d
- Standardized regression coefficient (beta; β)
- Correlation coefficient (r)
- Cohen's ω (omega)
- Cohen's f
- Cohen's f^2
- Coefficient of determination (R^2)
- Eta squared (η^2)
- Partial eta squared (η_p^2)

7.4.2.1 Cohen's d

Cohen's d is calculated as in Equation 7.12:

$$\begin{aligned} d &= \frac{\text{mean difference}}{\text{pooled standard deviation}} \\ &= \frac{\bar{X}_1 - \bar{X}_2}{s} \end{aligned} \tag{7.12}$$

where:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{7.13}$$

where n_1 and n_2 is the sample size of group 1 and group 2, respectively, and s_1 and s_2 is the standard deviation of group 1 and group 2, respectively.

7.4.2.2 Standardized Regression Coefficient (Beta; β)

The standardized regression coefficient (beta; β) is used in multiple regression, and is calculated as in Equation 7.14:

$$\beta_x = B_x \times \frac{s_x}{s_y} \tag{7.14}$$

where B_x is the unstandardized regression coefficient of the predictor variable x in predicting the outcome variable y, s_x is the standard deviation of x, and s_y is the standard deviation of y.

7.4.2.3 Correlation Coefficient (r)

The formula for the correlation coefficient is in Chapter 8.

7.4.2.4 Cohen's ω

Cohen's ω is used in chi-square tests, and is calculated as in Equation 7.15:

$$\omega = \sqrt{\frac{\chi^2}{N} - \frac{df}{N}} \tag{7.15}$$

where χ^2 is the chi-square statistic from the test, N is the sample size, and df is the degrees of freedom.

7.4.2.5 Cohen's f

Cohen's f is commonly used in ANOVA, and is calculated as in Equation 7.16:

$$\begin{aligned} f &= \sqrt{\frac{R^2}{1-R^2}} \\ &= \sqrt{\frac{\eta^2}{1-\eta^2}} \end{aligned} \tag{7.16}$$

7.4.2.6 Cohen's f^2

Cohen's f^2 is commonly used in regression, and is calculated as in Equation 7.17:

$$\begin{aligned} f^2 &= \frac{R^2}{1-R^2} \\ &= \frac{\eta^2}{1-\eta^2} \end{aligned} \tag{7.17}$$

To calculate the effect size of a particular predictor, you can calculate Δf^2 as in Equation 7.18:

$$\begin{aligned} \Delta f^2 &= \frac{R^2_{\text{model}} - R^2_{\text{reduced}}}{1 - R^2_{\text{model}}} \\ &= \frac{\eta^2_{\text{model}} - \eta^2_{\text{reduced}}}{1 - \eta^2_{\text{model}}} \end{aligned} \tag{7.18}$$

where R^2_{model} is the R^2 of the model with the predictor variable of interest and R^2_{reduced} is the R^2 of the model without the predictor variable of interest.

7.4.2.7 Coefficient of Determination (R^2)

The coefficient of determination (R^2) reflects the proportion of variance in the outcome variable that is explained by the predictor variable(s). R^2 is commonly used in regression, and is calculated as in Equation 7.19:

$$\begin{aligned} R^2 &= 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}} \\ &= 1 - \frac{\text{sum of squared residuals}}{\text{total sum of squares}} \\ &= \frac{f^2}{1+f^2} \\ &= \eta^2 \\ &= \frac{\text{variance explained in } Y}{\text{total variance in } Y} \end{aligned} \tag{7.19}$$

where Y_i is the observed value of the outcome variable for the ith observation, $\hat{Y}_i$ is the model predicted value for the ith observation, $\bar{Y}$ is the mean of the observed values of the outcome variable. The total sum of squares is an index of the total variation in the outcome variable.

7.4.2.8 Eta Squared (η^2) and Partial Eta Squared (η_p^2)

Like R^2, eta squared (η^2) reflects the proportion of variance in the dependent variable that is explained by the independent variable(s). η^2 is commonly used in ANOVA, and is calculated as in Equation 7.20:

$$\begin{aligned} \eta^2 &= \frac{SS_{\text{effect}}}{SS_{\text{total}}} \\ &= 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}} \\ &= 1 - \frac{\text{sum of squared residuals}}{\text{total sum of squares}} \\ &= \frac{f^2}{1+f^2} \\ &= R^2 \end{aligned} \tag{7.20}$$

where SS_{effect} is the sum of squares for the effect of interest and SS_{total} is the total sum of squares.

Partial eta squared (η_p^2) reflects the proportion of variance in the dependent variable that is explained by the independent variable while controlling for

the other independent variables. η_p^2 is commonly used in ANOVA, and is calculated as in Equation 7.21:

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}} \tag{7.21}$$

where SS_{effect} is the sum of squares for the effect of interest and SS_{error} is the sum of squares for the residual error term.

7.4.2.9 Effect Size Thresholds

Effect size thresholds (Cohen, 1988; McGrath & Meyer, 2006) for small, medium, and large effect sizes are in Table 7.2.

Table 7.2 Effect Size Thresholds for Small, Medium, and Large Effect Sizes.

Effect Size Index	Small	Medium	Large
Cohen's d	$\geq \lvert.20\rvert$	$\geq \lvert.50\rvert$	$\geq \lvert.80\rvert$
Standardized regression coefficient (beta; β)	$\geq \lvert.10\rvert$	$\geq \lvert.24\rvert$	$\geq \lvert.37\rvert$
Correlation coefficient (r)	$\geq \lvert.10\rvert$	$\geq \lvert.24\rvert$	$\geq \lvert.37\rvert$
Cohen's ω	$\geq .10$	$\geq .30$	$\geq .50$
Cohen's f	$\geq .10$	$\geq .25$	$\geq .40$
Cohen's f^2	$\geq .01$	$\geq .06$	$\geq .16$
Coefficient of determination (R^2)	$\geq .01$	$\geq .06$	$\geq .14$
Eta squared (η^2)	$\geq .01$	$\geq .06$	$\geq .14$
Partial eta squared (η_p^2)	$\geq .01$	$\geq .06$	$\geq .14$

7.5 Statistical Decision Tree

An example statistical decision tree is depicted in Figure 7.8.

However, many statistical tests can be re-formulated in a regression framework, as in Figure 7.9.

For an online, interactive statistical decision tree to help you decide which statistical analysis to use, see here: https://www.statsflowchart.co.uk

Figure 7.8 A Statistical Decision Tree For Choosing an Appropriate Statistical Procedure. Adapted from: https://commons.wikimedia.org/wiki/File:InferentialStatisticalDecisionMakingTrees.pdf. The original source is: Corston, R. & Colman, A. M. (2000). *A crash course in SPSS for Windows*. Wiley-Blackwell. Changes were made to the original, including the addition of several statistical tests. *Note*: "Interval" as a level of measurement includes data with an "interval" or higher level of measurement; thus, it also includes data with a "ratio" level of measurement.

7.6 Statistical Tests

7.6.1 *t*-Test

There are several *t*-tests:

- one-sample *t*-test
- two-samples *t*-test
  - independent samples *t*-test
  - paired samples *t*-test

Figure 7.9 A Statistical Decision Tree For Choosing an Appropriate Statistical Procedure, Re-Formulated in a Regression Framework. Adapted from: https://commons.wikimedia.org/wiki/File:InferentialStatisticalDecisionMakingTrees.pdf. The original source is: Corston, R. & Colman, A. M. (2000). *A crash course in SPSS for Windows*. Wiley-Blackwell. Changes were made to the original, including re-formulating the tests in a regression framework.

A one-sample t-test is used to evaluate whether a sample mean differs systematically from a particular value. The null hypothesis is that the sample mean does not differ systematically from the pre-specified value. The alternative hypothesis is that the sample mean differs systematically from the pre-specified value. For instance, let's say you want to test out a new draft strategy. You could participate in a mock draft and draft players using the new strategy. Then, you could use a one-sample t-test to evaluate whether your new draft strategy yields players with more projected points than the average of players' projected points for other teams.

Two-samples t-tests are used to test for differences between scores of two groups. If the two groups are independent, the independent samples t-test is

used. If the two groups involve paired samples, the paired samples t-test is used. The null hypothesis is that the mean of group 1 does not differ systematically from the mean of group 2. The alternative hypothesis is that the mean of group 1 differs systematically from the mean of group 2. For instance, you could use an independent-samples t-test if you want to examine whether Quarterbacks tend to have have longer careers than Running Backs. By contrast, you could use a paired samples t-test if you want to examine whether Quarterbacks tend to score more points in the second year of their contract compared to their rookie year, because the same subjects were assessed twice (i.e., a within-subject design).

7.6.2 Analysis of Variance

Analysis of variance (ANOVA) allows examining whether groups differ systematically as a function of one or more factors. There are multiple variants of ANOVA:

- one-way ANOVA
- factorial ANOVA
- repeated measures ANOVA (RM-ANOVA)
- multivariate ANOVA (MANOVA)

Like two-samples t-tests, ANOVA allows examining whether groups differ as a function of an independent variable. However, unlike a t-test, ANOVA allows examining multiple multiple independent variables and more than two groups. The null hypothesis is that the the groups' mean value does not differ systematically. The alternative hypothesis is that the groups' mean value differs systematically.

A one-way ANOVA examines whether two or more groups differ as a function of an independent variable. For instance, you could use a one-way ANOVA to evaluate if you want to evaluate whether multiple positions differ in their length of career. Factorial ANOVA examines whether two or more groups differ as a function of multiple independent variables. For instance, you could use factorial ANOVA to evaluate whether one's length of career depends on one's position and weight. Repeated measures ANOVA examines whether scores differ across repeated measures (e.g., across time) for the same participants. For instance, you could use repeated-measures ANOVA to evaluate whether rookies score more points as the season progresses. Multivariate ANOVA examines whether multiple dependent variables differ as a function of one or more factor(s). For instance, you could use MANOVA to evaluate whether one's contract length and pay differ as a function of one's position.

7.6.3 Correlation

Correlation examines the association between a predictor and outcome variable. The null hypothesis is that the the two variables are not associated. The alternative hypothesis is that the two variables are associated.

The Pearson correlation coefficient (r) is calculated as in Equation 7.22:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{7.22}$$

where X is the predictor variable and Y is the outcome variable.

7.6.4 (Multiple) Regression

Regression, like correlation, examines the association between a predictor and outcome variable. However, unlike correlation, regression allows multiple predictor variables.

Regression with a single predictor takes the form in Equation 9.1. A regression line is depicted in Figure 9.4. Multiple regression (i.e., regression with multiple predictors) takes the form in Equation 9.2.

The null hypothesis is that the the predictor variable(s) are not associated with the outcome variable. The alternative hypothesis is that the predictor variable(s) are associated with the outcome variable.

7.6.5 Chi-Square Test

There are two primary types of chi-square tests:

- chi-square goodness-of-fit test
- chi-square test for association (aka test of independence)

The chi-square goodness-of-fit test evaluates whether a set of categorical data came from a specified distribution. The null hypothesis is that the data came from the specified distribution. The alternative hypothesis is that the data did not come from the specified distribution.

The chi-square test for association evaluates whether two categorical variables are associated. The null hypothesis is that the two variables are not associated. The alternative hypothesis is that the two variables are associated.

7.6.6 Formulating Statistical Tests in Terms of Partitioned Variance

Many statistical tests can be formulated in terms of partitioned variance.

For instance, the t statistic from the independent-samples t-test and the F statistic from ANOVA can be thought of as the ratio of between-group variance to within-group variance, as in Equation 7.23:

$$t \text{ or } F = \frac{\text{between-group variance}}{\text{within-group variance}} \tag{7.23}$$

The correlation coefficient can be thought of as the ratio of shared variance (i.e., covariance) to total variance, as in Equation 7.24:

$$r = \frac{\text{shared variance}}{\text{total variance}} \tag{7.24}$$

The coefficient of determination (R^2) is the proportion of variance in the outcome variable that is explained by the predictor variables. η^2 is the proportion of variance in the dependent variable that is explained by the independent variables. The coefficient of determination and η^2 can be expressed as the ratio of variance explained in the outcome or dependent variable to the total variance in the outcome or dependent variable, as in Equation 7.25:

$$R^2 \text{ or } \eta^2 = \frac{\text{variance explained in the outcome variable}}{\text{total variance in the outcome variable}} \tag{7.25}$$

7.6.7 Critical Value

The critical value is the test value for a given test, above which the effect is considered to be statistically significant. The critical value for statistical significance for each test can be determined based on the degrees of freedom and alpha level. The degrees of freedom (df) refer to the number of values in the calculation of a test statistic that are free to vary.

```
alpha <- .05
N <- 200
nGroup1 <- 150
nGroup2 <- 150
numGroups <- 4
numLevelsFactorA <- 3
numLevelsFactorB <- 4
numMeasurements <- 4
numPredictors <- 5
numCategories <- 6
```

```
numRows <- 5
numColumns <- 2
```

7.6.7.1 One-Sample *t*-Test

For a one-sample t-test, the degrees of freedom is in Equation 7.26:

$$df = N - 1 \tag{7.26}$$

where N is sample size.

```
df_oneSampleTtest <- N - 1
```

One-tailed test:

```
qt(1 - alpha, df_oneSampleTtest)
```

```
[1] 1.652547
```

Two-tailed test:

```
qt(1 - alpha/2, df_oneSampleTtest)
```

```
[1] 1.971957
```

7.6.7.2 Independent-Samples *t*-Test

For an independent-samples t-test, the degrees of freedom is in Equation 7.27:

$$df = n_1 + n_2 - 2 \tag{7.27}$$

where n_1 is the sample size of group 1 and n_2 is the sample size of group 2.

```
df_independentSamplesTtest <- nGroup1 + nGroup2 - 2
```

One-tailed test:

```
qt(1 - alpha, df_independentSamplesTtest)
```

```
[1] 1.649983
```

Two-tailed test:

```
qt(1 - alpha/2, df_independentSamplesTtest)
```

```
[1] 1.967957
```

7.6.7.3 Paired-Samples *t*-Test

For a paired-samples t-test, the degrees of freedom is in Equation 7.28:

$$df = N - 1 \tag{7.28}$$

where N is sample size (i.e., the number of paired observations).

```
df_pairedSamplesTtest <- N - 1
```

One-tailed test:

```
qt(1 - alpha, df_pairedSamplesTtest)
```

```
[1] 1.652547
```

Two-tailed test:

```
qt(1 - alpha/2, df_pairedSamplesTtest)
```

```
[1] 1.971957
```

7.6.7.4 One-Way ANOVA

For a one-way ANOVA, the degrees of freedom is in Equation 7.29:

$$\begin{aligned} df_{\text{between}} &= g - 1 \\ df_{\text{within}} &= N - g \end{aligned} \tag{7.29}$$

where N is sample size and g is the number of groups.

```
df_betweenOneWayANOVA <- numGroups - 1
df_withinOneWayANOVA <- N - numGroups
```

One-tailed test:

```
qf(1 - alpha, df_betweenOneWayANOVA, df_withinOneWayANOVA)
```

```
[1] 2.650677
```

Two-tailed test:

```
qf(1 - alpha/2, df_betweenOneWayANOVA, df_withinOneWayANOVA)
```

```
[1] 3.183378
```

7.6.7.5 Factorial ANOVA

For a factorial two-way ANOVA, the degrees of freedom is in Equation 7.30:

$$\begin{aligned} df_{\text{Factor A}} &= a - 1 \\ df_{\text{Factor B}} &= b - 1 \\ df_{\text{Interaction}} &= (a-1)(b-1) \\ df_{\text{error}} &= ab(N-1) \end{aligned} \tag{7.30}$$

where N is sample size, a is the number of levels for factor A, and b is the number of levels for factor B.

```
df_factorA <- numLevelsFactorA - 1
df_factorB <- numLevelsFactorB - 1
df_interaction <- df_factorA * df_factorB
df_error <- numLevelsFactorA * numLevelsFactorB * (N - 1)
```

Factor A (one-tailed test):

```
qf(1 - alpha, df_factorA, df_error)
```

```
[1] 2.999494
```

Factor B (one-tailed test):

```
qf(1 - alpha, df_factorB, df_error)
```

```
[1] 2.608629
```

Interaction (one-tailed test):

```
qf(1 - alpha, df_interaction, df_error)
```

```
[1] 2.102376
```

Factor A (two-tailed test):

```
qf(1 - alpha/2, df_factorA, df_error)
```

```
[1] 3.694584
```

Factor B (two-tailed test):

```
qf(1 - alpha/2, df_factorB, df_error)
```

```
[1] 3.121587
```

Interaction (two-tailed test):

```
qf(1 - alpha/2, df_interaction, df_error)
```

```
[1] 2.413504
```

7.6.7.6 Repeated Measures ANOVA

For a repeated measures ANOVA, the degrees of freedom is in Equation 7.31:

$$\begin{aligned} df_1 &= T - 1 \\ df_2 &= (T-1)(N-1) \end{aligned} \tag{7.31}$$

where N is sample size and T is the number of measurements (i.e., the number of levels of the within-person factor: e.g., timepoints or conditions).

```
df1_RMANOVA <- numMeasurements - 1
df2_RMANOVA <- (numMeasurements - 1) * (N - 1)
```

One-tailed test:

```
qf(1 - alpha, df1_RMANOVA, df2_RMANOVA)
```

```
[1] 2.619828
```

Two-tailed test:

```
qf(1 - alpha/2, df1_RMANOVA, df2_RMANOVA)
```

```
[1] 3.138017
```

7.6.7.7 Correlation

For a correlation, the degrees of freedom is in Equation 7.32:

$$df = N - 2 \tag{7.32}$$

where N is sample size.

```
df_correlation <- N - 2
```

One-tailed test:

```
qt(1 - alpha, df_correlation)
```

```
[1] 1.652586
```

Two-tailed test:

```
qt(1 - alpha/2, df_correlation)
```

```
[1] 1.972017
```

7.6.7.8 Multiple Regression

For multiple regression, the degrees of freedom is in Equation 7.33:

$$\begin{aligned} df_1 &= p \\ df_2 &= N - p - 1 \end{aligned} \tag{7.33}$$

where N is sample size and p is the number of predictors.

```
df1_regression <- numPredictors
df2_regression <- N - numPredictors - 1
```

One-tailed test:

```
qf(1 - alpha, df1_regression, df2_regression)
```

```
[1] 2.260647
```

Two-tailed test:

```
qf(1 - alpha/2, df1_regression, df2_regression)
```

```
[1] 2.63243
```

7.6.7.9 Chi-Square Goodness-of-Fit Test

For the chi-square goodness-of-fit test, the degrees of freedom is in Equation 7.34:

$$df = c - 1 \tag{7.34}$$

where c is the number of categories.

```
df_chisquareGOF <- numCategories - 1
```

One-tailed test:

```
qchisq(1 - alpha, df_chisquareGOF)
```

```
[1] 11.0705
```

Two-tailed test:

```
qchisq(1 - alpha/2, df_chisquareGOF)
```

```
[1] 12.8325
```

7.6.7.10 Chi-Square Test for Association

For the chi-square test for association, the degrees of freedom is in Equation 7.35:

$$df = (r - 1) \times (c - 1) \tag{7.35}$$

where r is the number of rows in the contingency table and c is the number of columns in the contingency table.

```
df_chisquareAssociation <- (numRows - 1) * (numColumns - 1)
```

One-tailed test:

```
qchisq(1 - alpha, df_chisquareAssociation)
```

```
[1] 9.487729
```

Two-tailed test:

```
qchisq(1 - alpha/2, df_chisquareAssociation)
```

```
[1] 11.14329
```

7.6.8 Statistical Power

As described above, *statistical power* is the probability of detecting an effect, if, in fact, the effect exists. Statistical power for a given test can be calculated based on three factors:

- effect size
- sample size
- alpha level

Knowing any three of the following, you can calculate the fourth: statistical power, effect size, sample size, and alpha level. Below is `R` code for calculating power for each of various statistical tests (i.e., a *power analysis*). For free point-and-click software for calculating statistical power, see G*Power: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

```
power <- .8
effectSize_d <- .5
effectSize_r <- .24
effectSize_beta <- .24
effectSize_f <- .25
effectSize_fSquared <- .06
effectSize_omega <- .3
```

When designing a study, it is important to consider power and the sample size needed to detect the hypothesized effect size. If your sample size is too small and you do not detect an effect (i.e., $p > .05$), you do not know whether your failure to detect the effect was because a) the effect does not exist, or b) the effect exists but you did not have enough power to detect it.

7.6.8.1 One-Sample *t*-Test

Solving for statistical power achieved (given effect size, sample size, and alpha level):

```
pwr::pwr.t.test(
  n = N,
  d = effectSize_d,
  sig.level = alpha,
  type = "one.sample",
  alternative = "two.sided")
```

```
     One-sample t test power calculation

              n = 200
              d = 0.5
      sig.level = 0.05
          power = 0.9999998
    alternative = two.sided
```

Solving for sample size needed (given effect size, power, and alpha level):

```
pwr::pwr.t.test(
  power = power,
  d = effectSize_d,
  sig.level = alpha,
  type = "one.sample",
  alternative = "two.sided")
```

```
     One-sample t test power calculation

              n = 33.36713
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Solving for the minimum detectable effect size (given sample size, power, and alpha level):

```r
pwr::pwr.t.test(
  power = power,
  n = N,
  sig.level = alpha,
  type = "one.sample",
  alternative = "two.sided")
```

```
     One-sample t test power calculation

              n = 200
              d = 0.1990655
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

7.6.8.2 Independent-Samples *t*-Test

7.6.8.2.1 Balanced Group Sizes

Solving for statistical power achieved (given effect size, sample size per group, and alpha level):

```r
pwr::pwr.t.test(
  n = N,
  d = effectSize_d,
  sig.level = alpha,
  type = "two.sample",
  alternative = "two.sided")
```

```
     Two-sample t test power calculation

              n = 200
              d = 0.5
      sig.level = 0.05
          power = 0.9987689
    alternative = two.sided

NOTE: n is number in *each* group
```

Solving for sample size per group needed (given effect size, power, and alpha level):

```r
pwr::pwr.t.test(
  power = power,
  d = effectSize_d,
  sig.level = alpha,
  type = "two.sample",
  alternative = "two.sided")
```

```
     Two-sample t test power calculation

              n = 63.76561
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Solving for the minimum detectable effect size (given sample size per group, power, and alpha level):

```r
pwr::pwr.t.test(
  power = power,
  n = N,
  sig.level = alpha,
  type = "two.sample",
  alternative = "two.sided")
```

```
     Two-sample t test power calculation

              n = 200
              d = 0.2808267
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

7.6.8.2.2 Unbalanced Group Sizes

Solving for statistical power achieved (given effect size, sample size per group, and alpha level):

```r
pwr::pwr.t2n.test(
  n1 = nGroup1,
  n2 = nGroup2,
  d = effectSize_d,
  sig.level = alpha,
  alternative = "two.sided")
```

```
     t test power calculation

             n1 = 150
             n2 = 150
              d = 0.5
      sig.level = 0.05
          power = 0.9907677
    alternative = two.sided
```

Solving for sample size per group needed (given effect size, power, and alpha level):

```r
pwr::pwr.t2n.test(
  power = power,
  n1 = nGroup1,
  d = effectSize_d,
  sig.level = alpha,
  alternative = "two.sided")
```

```
     t test power calculation

             n1 = 150
             n2 = 40.22483
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Solving for the minimum detectable effect size (given sample size per group, power, and alpha level):

```r
pwr::pwr.t2n.test(
  power = power,
  n1 = nGroup1,
```

```
  n2 = nGroup2,
  sig.level = alpha,
  alternative = "two.sided")
```

```
     t test power calculation

             n1 = 150
             n2 = 150
              d = 0.3245459
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

7.6.8.3 Paired-Samples *t*-Test

Solving for statistical power achieved (given effect size, sample size per group, and alpha level):

```
pwr::pwr.t.test(
  n = N,
  d = effectSize_d,
  sig.level = alpha,
  type = "paired",
  alternative = "two.sided")
```

```
     Paired t test power calculation

              n = 200
              d = 0.5
      sig.level = 0.05
          power = 0.9999998
    alternative = two.sided

NOTE: n is number of *pairs*
```

Solving for sample size per group needed (given effect size, power, and alpha level):

```
pwr::pwr.t.test(
  power = power,
  d = effectSize_d,
```

```
  sig.level = alpha,
  type = "paired",
  alternative = "two.sided")
```

```
     Paired t test power calculation

              n = 33.36713
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

Solving for the minimum detectable effect size (given sample size per group, power, and alpha level):

```
pwr::pwr.t.test(
  power = power,
  n = N,
  sig.level = alpha,
  type = "paired",
  alternative = "two.sided")
```

```
     Paired t test power calculation

              n = 200
              d = 0.1990655
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

7.6.8.4 One-Way ANOVA

Solving for statistical power achieved (given effect size, sample size per group, and alpha level):

```r
pwr::pwr.anova.test(
  n = N,
  f = effectSize_f,
  sig.level = alpha,
  k = numGroups)
```

```
     Balanced one-way analysis of variance power calculation

              k = 4
              n = 200
              f = 0.25
      sig.level = 0.05
          power = 0.9999962

NOTE: n is number in each group
```

Solving for sample size per group needed (given effect size, power, and alpha level):

```r
pwr::pwr.anova.test(
  power = power,
  f = effectSize_f,
  sig.level = alpha,
  k = numGroups)
```

```
     Balanced one-way analysis of variance power calculation

              k = 4
              n = 44.59927
              f = 0.25
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

Solving for the minimum detectable effect size (given sample size per group, power, and alpha level):

```r
pwr::pwr.anova.test(
  power = power,
  n = N,
  sig.level = alpha,
  k = numGroups)
```

```
     Balanced one-way analysis of variance power calculation

              k = 4
              n = 200
              f = 0.117038
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

The power analysis code above assumes the groups are of equal size (i.e., a balanced design). If the design is unbalanced (i.e., there are different numbers of participants in each group), it may be necessary to conduct a power analysis via a simulation. Below is an example of evaluating the statistical power for detecting an effect unbalanced designs via simulation:

```r
nSim <- 1000 # number of simulations

# Function to generate data and perform ANOVA
simulate_anova <- function(nGroup1, nGroup2, f, alpha) {
  # Means for each group
  mean1 <- 0
  mean2 <- f * sqrt((nGroup1 + nGroup2) / 2)

  # Generate data
  group1 <- rnorm(nGroup1, mean = mean1, sd = 1)
  group2 <- rnorm(nGroup2, mean = mean2, sd = 1)

  # Combine data
  data <- data.frame(
    value = c(group1, group2),
    group = factor(rep(c("Group1", "Group2"), c(nGroup1, nGroup2)))
  )

  # Perform ANOVA
  aov_result <- aov(value ~ group, data = data)
  p_value <- summary(aov_result)[[1]][["Pr(>F)"]][1]

  # Check if p-value is less than alpha
  return(p_value < alpha)
}

# Run simulations
set.seed(52242) # for reproducibility
```

```r
powerSimulationOneWayAnova <- replicate(
  nSim,
  simulate_anova(
    nGroup1 = 10,
    nGroup2 = 25,
    f = effectSize_f,
    alpha = alpha))

# Estimate power
mean(powerSimulationOneWayAnova)
```

```
[1] 0.774
```

7.6.8.5 Factorial ANOVA

The power analysis code below assumes the groups are of equal size (i.e., a balanced design). If the design is unbalanced (i.e., there are different numbers of participants in each group), it may be necessary to conduct a power analysis via a simulation. See Section 7.6.8.4 for an example power analysis simulation for one-way ANOVA.

Solving for statistical power achieved (given effect size, sample size per group, and alpha level):

```r
pwr::pwr.anova.test(
  n = N,
  f = effectSize_f,
  sig.level = alpha,
  k = numLevelsFactorA)
```

```
     Balanced one-way analysis of variance power calculation

              k = 3
              n = 200
              f = 0.25
      sig.level = 0.05
          power = 0.9999238

NOTE: n is number in each group
```

```r
pwr::pwr.anova.test(
  n = N,
  f = effectSize_f,
```

```
  sig.level = alpha,
  k = numLevelsFactorB)
```

```
     Balanced one-way analysis of variance power calculation

              k = 4
              n = 200
              f = 0.25
      sig.level = 0.05
          power = 0.9999962

NOTE: n is number in each group
```

```
pwr::pwr.anova.test(
  n = N,
  f = effectSize_f,
  sig.level = alpha,
  k = numLevelsFactorA + numLevelsFactorB)
```

```
     Balanced one-way analysis of variance power calculation

              k = 7
              n = 200
              f = 0.25
      sig.level = 0.05
          power = 1

NOTE: n is number in each group
```

Solving for sample size per group needed (given effect size, power, and alpha level):

```
pwr::pwr.anova.test(
  power = power,
  f = effectSize_f,
  sig.level = alpha,
  k = numLevelsFactorA)
```

```
     Balanced one-way analysis of variance power calculation
```

```
              k = 3
              n = 52.3966
              f = 0.25
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

```
pwr::pwr.anova.test(
  power = power,
  f = effectSize_f,
  sig.level = alpha,
  k = numLevelsFactorB)
```

```
     Balanced one-way analysis of variance power calculation

              k = 4
              n = 44.59927
              f = 0.25
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

```
pwr::pwr.anova.test(
  power = power,
  f = effectSize_f,
  sig.level = alpha,
  k = numLevelsFactorA + numLevelsFactorB)
```

```
     Balanced one-way analysis of variance power calculation

              k = 7
              n = 32.05196
              f = 0.25
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

Solving for the minimum detectable effect size (given sample size per group, power, and alpha level):

```r
pwr::pwr.anova.test(
  power = power,
  n = N,
  sig.level = alpha,
  k = numLevelsFactorA)
```

```
     Balanced one-way analysis of variance power calculation

              k = 3
              n = 200
              f = 0.1270373
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

```r
pwr::pwr.anova.test(
  power = power,
  n = N,
  sig.level = alpha,
  k = numLevelsFactorB)
```

```
     Balanced one-way analysis of variance power calculation

              k = 4
              n = 200
              f = 0.117038
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

```r
pwr::pwr.anova.test(
  power = power,
  n = N,
  sig.level = alpha,
  k = numLevelsFactorA + numLevelsFactorB)
```

```
     Balanced one-way analysis of variance power calculation
```

```
              k = 7
              n = 200
              f = 0.09889082
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

7.6.8.6 Repeated Measures ANOVA

Solving for statistical power achieved (given effect size, sample size per group, and alpha level):

```
WebPower::wp.rmanova(
  n = N,
  ng = numGroups,
  nm = numMeasurements,
  f = effectSize_f,
  alpha = alpha,
  type = 0)
```

```
Repeated-measures ANOVA analysis

      n    f ng nm nscor alpha     power
    200 0.25  4  4     1  0.05 0.8484718

NOTE: Power analysis for between-effect test
URL: http://psychstat.org/rmanova
```

```
WebPower::wp.rmanova(
  n = N,
  ng = numGroups,
  nm = numMeasurements,
  f = effectSize_f,
  alpha = alpha,
  type = 1)
```

```
Repeated-measures ANOVA analysis

      n    f ng nm nscor alpha     power
    200 0.25  4  4     1  0.05 0.8536292

NOTE: Power analysis for within-effect test
URL: http://psychstat.org/rmanova
```

```r
WebPower::wp.rmanova(
  n = N,
  ng = numGroups,
  nm = numMeasurements,
  f = effectSize_f,
  alpha = alpha,
  type = 2)
```

```
Repeated-measures ANOVA analysis

      n    f ng nm nscor alpha     power
    200 0.25  4  4     1  0.05 0.6756298

NOTE: Power analysis for interaction-effect test
URL: http://psychstat.org/rmanova
```

Solving for sample size per group needed (given effect size, power, and alpha level):

```r
WebPower::wp.rmanova(
  power = power,
  ng = numGroups,
  nm = numMeasurements,
  f = effectSize_f,
  alpha = alpha,
  type = 0)
```

```
Repeated-measures ANOVA analysis

           n    f ng nm nscor alpha power
    178.3971 0.25  4  4     1  0.05   0.8

NOTE: Power analysis for between-effect test
URL: http://psychstat.org/rmanova
```

```r
WebPower::wp.rmanova(
  power = power,
  ng = numGroups,
  nm = numMeasurements,
  f = effectSize_f,
  alpha = alpha,
  type = 1)
```

```
Repeated-measures ANOVA analysis

           n    f ng nm nscor alpha power
    175.7692 0.25  4  4     1  0.05   0.8

NOTE: Power analysis for within-effect test
URL: http://psychstat.org/rmanova
```

```
WebPower::wp.rmanova(
  power = power,
  ng = numGroups,
  nm = numMeasurements,
  f = effectSize_f,
  alpha = alpha,
  type = 2)
```

```
Repeated-measures ANOVA analysis

           n    f ng nm nscor alpha power
    253.2369 0.25  4  4     1  0.05   0.8

NOTE: Power analysis for interaction-effect test
URL: http://psychstat.org/rmanova
```

Solving for the minimum detectable effect size (given sample size per group, power, and alpha level):

```
WebPower::wp.rmanova(
  power = power,
  n = N,
  ng = numGroups,
  nm = numMeasurements,
  alpha = alpha,
  type = 0)
```

```
Repeated-measures ANOVA analysis

      n         f ng nm nscor alpha power
    200 0.2358259  4  4     1  0.05   0.8

NOTE: Power analysis for between-effect test
URL: http://psychstat.org/rmanova
```

```r
WebPower::wp.rmanova(
  power = power,
  n = N,
  ng = numGroups,
  nm = numMeasurements,
  alpha = alpha,
  type = 1)
```

```
Repeated-measures ANOVA analysis

      n         f ng nm nscor alpha power
    200 0.2342726  4  4     1  0.05   0.8

NOTE: Power analysis for within-effect test
URL: http://psychstat.org/rmanova
```

```r
WebPower::wp.rmanova(
  power = power,
  n = N,
  ng = numGroups,
  nm = numMeasurements,
  alpha = alpha,
  type = 2)
```

```
Repeated-measures ANOVA analysis

      n         f ng nm nscor alpha power
    200 0.2817486  4  4     1  0.05   0.8

NOTE: Power analysis for interaction-effect test
URL: http://psychstat.org/rmanova
```

7.6.8.7 Correlation

Solving for statistical power achieved (given effect size, sample size per group, and alpha level):

```r
pwr::pwr.r.test(
  n = N,
  r = effectSize_r,
  sig.level = alpha,
  alternative = "two.sided")
```

```
     approximate correlation power calculation (arctangh transformation)

              n = 200
              r = 0.24
      sig.level = 0.05
          power = 0.9310138
    alternative = two.sided
```

Solving for sample size per group needed (given effect size, power, and alpha level):

```
pwr::pwr.r.test(
  power = power,
  r = effectSize_r,
  sig.level = alpha,
  alternative = "two.sided")
```

```
     approximate correlation power calculation (arctangh transformation)

              n = 133.1299
              r = 0.24
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Solving for the minimum detectable effect size (given sample size per group, power, and alpha level):

```
pwr::pwr.r.test(
  power = power,
  n = N,
  sig.level = alpha,
  alternative = "two.sided")
```

```
     approximate correlation power calculation (arctangh transformation)

              n = 200
              r = 0.1965767
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

7.6.8.8 Multiple Regression

Solving for statistical power achieved (given effect size, sample size, and alpha level):

```
pwr::pwr.f2.test(
  f2 = effectSize_fSquared,
  sig.level = alpha,
  u = numPredictors,
  v = N - numPredictors - 1)
```

```
     Multiple regression power calculation

              u = 5
              v = 194
             f2 = 0.06
      sig.level = 0.05
          power = 0.7548031
```

```
pwrss::pwrss.t.reg(
  n = N,
  beta1 = effectSize_beta,
  k = numPredictors,
  alpha = alpha,
  alternative = "not equal")
```

```
 Linear Regression Coefficient (t Test)
 H0: beta1 = beta0
 HA: beta1 != beta0
 ------------------------------
  Statistical power = 0.936
  n = 200
 ------------------------------
 Alternative = "not equal"
 Degrees of freedom = 194
 Non-centrality parameter = 3.496
 Type I error rate = 0.05
 Type II error rate = 0.064
```

Solving for sample size needed (given effect size, power, and alpha level)—$v = N - \text{numberOfPredictors} - 1$; thus, $N = v + \text{numberOfPredictors} + 1$:

```r
multipleRegressionSampleSizeModel <- pwr::pwr.f2.test(
  power = power,
  f2 = effectSize_fSquared,
  sig.level = alpha,
  u = numPredictors)

multipleRegressionSampleSizeModel
```

```
     Multiple regression power calculation

              u = 5
              v = 213.3947
             f2 = 0.06
      sig.level = 0.05
          power = 0.8
```

```r
vNeeded <- multipleRegressionSampleSizeModel$v
sampleSizeNeeded <- vNeeded + numPredictors + 1
sampleSizeNeeded
```

```
[1] 219.3947
```

```r
pwrss::pwrss.t.reg(
  power = power,
  beta1 = effectSize_beta,
  k = numPredictors,
  alpha = alpha,
  alternative = "not equal")
```

```
 Linear Regression Coefficient (t Test)
 H0: beta1 = beta0
 HA: beta1 != beta0
 ------------------------------
  Statistical power = 0.8
  n = 131
 ------------------------------
 Alternative = "not equal"
 Degrees of freedom = 124.427
 Non-centrality parameter = 2.823
 Type I error rate = 0.05
 Type II error rate = 0.2
```

Solving for the minimum detectable effect size (given sample size, power, and alpha level):

```
pwr::pwr.f2.test(
  power = power,
  sig.level = alpha,
  u = numPredictors,
  v = N - numPredictors - 1)
```

```
     Multiple regression power calculation

              u = 5
              v = 194
             f2 = 0.06597765
      sig.level = 0.05
          power = 0.8
```

7.6.8.9 Chi-Square Goodness-of-Fit Test

Solving for statistical power achieved (given effect size, sample size, and alpha level):

```
pwr::pwr.chisq.test(
  N = N,
  w = effectSize_omega,
  df = numCategories - 1,
  sig.level = alpha)
```

```
     Chi squared power calculation

              w = 0.3
              N = 200
             df = 5
      sig.level = 0.05
          power = 0.9269225

NOTE: N is the number of observations
```

Solving for sample size needed (given effect size, power, and alpha level):

```r
pwr::pwr.chisq.test(
  power = power,
  w = effectSize_omega,
  df = numCategories - 1,
  sig.level = alpha)
```

```
     Chi squared power calculation

              w = 0.3
              N = 142.529
             df = 5
      sig.level = 0.05
          power = 0.8

NOTE: N is the number of observations
```

Solving for the minimum detectable effect size (given sample size, power, and alpha level):

```r
pwr::pwr.chisq.test(
  power = power,
  N = N,
  df = numCategories - 1,
  sig.level = alpha)
```

```
     Chi squared power calculation

              w = 0.2532543
              N = 200
             df = 5
      sig.level = 0.05
          power = 0.8

NOTE: N is the number of observations
```

7.6.8.10 Chi-Square Test for Association

Solving for statistical power achieved (given effect size, sample size, and alpha level):

```r
pwr::pwr.chisq.test(
  N = N,
  w = effectSize_omega,
  df = (numRows - 1)*(numColumns - 1),
  sig.level = alpha)
```

```
     Chi squared power calculation

              w = 0.3
              N = 200
             df = 4
      sig.level = 0.05
          power = 0.9431195

NOTE: N is the number of observations
```

Solving for sample size needed (given effect size, power, and alpha level):

```r
pwr::pwr.chisq.test(
  power = power,
  w = effectSize_omega,
  df = (numRows - 1)*(numColumns - 1),
  sig.level = alpha)
```

```
     Chi squared power calculation

              w = 0.3
              N = 132.6143
             df = 4
      sig.level = 0.05
          power = 0.8

NOTE: N is the number of observations
```

Solving for the minimum detectable effect size (given sample size, power, and alpha level):

```r
pwr::pwr.chisq.test(
  power = power,
  N = N,
  df = (numRows - 1)*(numColumns - 1),
  sig.level = alpha)
```

```
     Chi squared power calculation

              w = 0.2442875
              N = 200
             df = 4
      sig.level = 0.05
          power = 0.8

NOTE: N is the number of observations
```

7.6.8.11 Multilevel Modeling

Power analysis for multilevel modeling approaches is more complicated than it is for other statistical analyses, such as correlation, multiple regression, *t*-tests, ANOVA[2], etc.

There are free web applications for calculating power in multilevel modeling:

- https://aguinis.shinyapps.io/ml_power/
- https://koumurayama.shinyapps.io/tmethod_mlm/
- https://webpower.psychstat.org/wiki/models/index

7.6.8.12 Path Analysis, Factor Analysis, and Structural Equation Modeling

Power analysis for latent variable modeling approaches like structural equation modeling (SEM) is more complicated than it is for other statistical analyses, such as correlation, multiple regression, *t*-tests, ANOVA[3], etc.

I provide an example of power analysis in SEM using Monte Carlo simulation in R here: https://isaactpetersen.github.io/Principles-Psychological-Assessment/sem.html#monteCarloPowerAnalysis (Petersen, 2024c).

There are also free web applications for calculating power in SEM:

- https://sjak.shinyapps.io/power4SEM/
- https://sempower.shinyapps.io/sempower/
- https://yilinandrewang.shinyapps.io/pwrSEM/
- https://webpower.psychstat.org/wiki/models/index

[2] @sec-anova
[3] @sec-anova

7.6.8.13 Mediation and Moderation

There are free tools for calculating power for tests of mediation and moderation:

- https://schoemanna.shinyapps.io/mc_power_med/
- https://www.causalevaluation.org/power-analysis.html (web application: https://powerupr.shinyapps.io/index/)
- https://webpower.psychstat.org/wiki/models/index

8

Correlation Analysis

8.1 Getting Started

8.1.1 Load Packages

```
library("petersenlab")
library("XICOR")
library("tidyverse")
```

8.2 Overview of Correlation

Correlation is an index of the association between variables. Covariance is the association between variables and in an unstandardized metric that differs for variables with different scales. By contrast, correlation is in a standarized metric that does not differ for variables with different scales. When examining the association between variables that are interval or ratio levels of measurement, Pearson correlation is used. When examining the association between variables that are ordinal in level of measurement, Spearman correlation is used. Pearson correlation is an index of the *linear* association between variables. If a nonlinear association is present, other indices like xi [ξ; Chatterjee (2021)] and distance correlation coefficients are better suited to detect the association.

8.3 The Correlation Coefficient (r)

The formula for the correlation coefficient is in Equation 7.22.

The correlation coefficient ranges from −1.0 to +1.0. The correlation coefficient (r) tells you two things: (1) the direction (sign) of the association (positive or negative) and (2) the magnitude of the association. If the correlation coefficient is positive, the association is positive. If the correlation coefficient is negative, the association is negative. If the association is positive, as `X` increases, `Y` increases (or conversely, as `X` decreases, `Y` decreases). If the association is negative, as `X` increases, `Y` decreases (or conversely, as `X` decreases, `Y` increases). The smaller the absolute value of the correlation coefficient (i.e., the closer the r value is to zero), the weaker the association and the flatter the slope of the best-fit line in a scatterplot. The larger the absolute value of the correlation coefficient (i.e., the closer the absolute value of the r value is to one), the stronger the association and the steeper the slope of the best-fit line in a scatterplot. See Figure 8.1 for a range of different correlation coefficients and what some example data may look like for each direction and strength of association.

```
set.seed(52242)
correlations <- data.frame(criterion = rnorm(1000))

correlations$v1 <- complement(correlations$criterion, -1)
correlations$v2 <- complement(correlations$criterion, -.9)
correlations$v3 <- complement(correlations$criterion, -.8)
correlations$v4 <- complement(correlations$criterion, -.7)
correlations$v5 <- complement(correlations$criterion, -.6)
correlations$v6 <- complement(correlations$criterion, -.5)
correlations$v7 <- complement(correlations$criterion, -.4)
correlations$v8 <- complement(correlations$criterion, -.3)
correlations$v9 <- complement(correlations$criterion, -.2)
correlations$v10 <-complement(correlations$criterion, -.1)
correlations$v11 <-complement(correlations$criterion, 0)
correlations$v12 <-complement(correlations$criterion, .1)
correlations$v13 <-complement(correlations$criterion, .2)
correlations$v14 <-complement(correlations$criterion, .3)
correlations$v15 <-complement(correlations$criterion, .4)
correlations$v16 <-complement(correlations$criterion, .5)
correlations$v17 <-complement(correlations$criterion, .6)
correlations$v18 <-complement(correlations$criterion, .7)
correlations$v19 <-complement(correlations$criterion, .8)
correlations$v20 <-complement(correlations$criterion, .9)
correlations$v21 <-complement(correlations$criterion, 1)

par(mfrow = c(7,3), mar = c(1, 0, 1, 0))

# -1.0
plot(correlations$criterion, correlations$v1, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
```

```
     main = substitute(paste(italic(r), " = ", x, sep = ""), list(x = round(cor.test(x = correlati
abline(lm(v1 ~ criterion, data = correlations), col = "black")

# -.9
plot(correlations$criterion, correlations$v2, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = ""), list(x = round(cor.test(x = correlati
abline(lm(v2 ~ criterion, data = correlations), col = "black")

# -.8
plot(correlations$criterion, correlations$v3, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = ""), list(x = round(cor.test(x = correlati
abline(lm(v3 ~ criterion, data = correlations), col = "black")

# -.7
plot(correlations$criterion, correlations$v4, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = ""), list(x = round(cor.test(x = correlati
abline(lm(v4 ~ criterion, data = correlations), col = "black")

# -.6
plot(correlations$criterion, correlations$v5, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = ""), list(x = round(cor.test(x = correlati
abline(lm(v5 ~ criterion, data = correlations), col = "black")

# -.5
plot(correlations$criterion, correlations$v6, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = ""), list(x = round(cor.test(x = correlati
abline(lm(v6 ~ criterion, data = correlations), col = "black")

# -.4
plot(correlations$criterion, correlations$v7, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = ""), list(x = round(cor.test(x = correlati
abline(lm(v7 ~ criterion, data = correlations), col = "black")

# -.3
plot(correlations$criterion, correlations$v8, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = ""), list(x = round(cor.test(x = correlati
abline(lm(v8 ~ criterion, data = correlations), col = "black")

# -.2
plot(correlations$criterion, correlations$v9, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = ""), list(x = round(cor.test(x = correlati
abline(lm(v9 ~ criterion, data = correlations), col = "black")

# -.1
```

```
plot(correlations$criterion, correlations$v10, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v10 ~ criterion, data = correlations), col = "black")

# 0.0
plot(correlations$criterion, correlations$v11, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v11 ~ criterion, data = correlations), col = "black")

# 0.1
plot(correlations$criterion, correlations$v12, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v12 ~ criterion, data = correlations), col = "black")

# 0.2
plot(correlations$criterion, correlations$v13, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v13 ~ criterion, data = correlations), col = "black")

# 0.3
plot(correlations$criterion, correlations$v14, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v14 ~ criterion, data = correlations), col = "black")

# 0.4
plot(correlations$criterion, correlations$v15, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v15 ~ criterion, data = correlations), col = "black")

# 0.5
plot(correlations$criterion, correlations$v16, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v16 ~ criterion, data = correlations), col = "black")

# 0.6
plot(correlations$criterion, correlations$v17, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v17 ~ criterion, data = correlations), col = "black")

# 0.7
plot(correlations$criterion, correlations$v18, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v18 ~ criterion, data = correlations), col = "black")
```

```
# 0.8
plot(correlations$criterion, correlations$v19, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v19 ~ criterion, data = correlations), col = "black")

# 0.9
plot(correlations$criterion, correlations$v20, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v20 ~ criterion, data = correlations), col = "black")

# 1.0
plot(correlations$criterion, correlations$v21, xaxt = "n", yaxt = "n", xlab = "" , ylab = "",
     main = substitute(paste(italic(r), " = ", x, sep = "")), list(x = round(cor.test(x = correlati
abline(lm(v21 ~ criterion, data = correlations), col = "black")

invisible(dev.off()) #par(mfrow = c(1,1))
```

Figure 8.1 Correlation Coefficients.

See Figure 8.2 for the interpretation of the magnitude and direction (sign) of various correlation coefficients.

```
library("patchwork")

set.seed(52242)
correlations2 <- data.frame(criterion = rnorm(15))

correlations2$v1 <- complement(correlations2$criterion, -1)
correlations2$v2 <- complement(correlations2$criterion, -.9)
correlations2$v3 <- complement(correlations2$criterion, -.8)
correlations2$v4 <- complement(correlations2$criterion, -.7)
correlations2$v5 <- complement(correlations2$criterion, -.6)
correlations2$v6 <- complement(correlations2$criterion, -.5)
correlations2$v7 <- complement(correlations2$criterion, -.4)
correlations2$v8 <- complement(correlations2$criterion, -.3)
correlations2$v9 <- complement(correlations2$criterion, -.2)
correlations2$v10 <-complement(correlations2$criterion, -.1)
correlations2$v11 <-complement(correlations2$criterion, 0)
correlations2$v12 <-complement(correlations2$criterion, .1)
correlations2$v13 <-complement(correlations2$criterion, .2)
correlations2$v14 <-complement(correlations2$criterion, .3)
correlations2$v15 <-complement(correlations2$criterion, .4)
correlations2$v16 <-complement(correlations2$criterion, .5)
correlations2$v17 <-complement(correlations2$criterion, .6)
correlations2$v18 <-complement(correlations2$criterion, .7)
correlations2$v19 <-complement(correlations2$criterion, .8)
correlations2$v20 <-complement(correlations2$criterion, .9)
correlations2$v21 <-complement(correlations2$criterion, 1)

# -1.0
p1 <- ggplot(
  data = correlations2,
  mapping = aes(
    x = criterion,
    y = v1
  )
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE) +
  labs(
    title = "Perfect Negative Association",
    subtitle = expression(paste(italic("r"), " = ", "-1.0"))
```

```
  ) +
  theme_classic(
    base_size = 12) +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())

# -0.9
p2 <- ggplot(
  data = correlations2,
  mapping = aes(
    x = criterion,
    y = v2
  )
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE) +
  labs(
    title = "Strong Negative Association",
    subtitle = expression(paste(italic("r"), " = ", "-.9"))
  ) +
  theme_classic(
    base_size = 12) +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())

# -0.5
p3 <- ggplot(
  data = correlations2,
  mapping = aes(
    x = criterion,
    y = v6
  )
```

```
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE) +
  labs(
    title = "Moderate Negative Association",
    subtitle = expression(paste(italic("r"), " = ", "-.5"))
  ) +
  theme_classic(
    base_size = 12) +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())

# -0.2
p4 <- ggplot(
  data = correlations2,
  mapping = aes(
    x = criterion,
    y = v9
  )
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE) +
  labs(
    title = "Weak Negative Association",
    subtitle = expression(paste(italic("r"), " = ", "-.2"))
  ) +
  theme_classic(
    base_size = 12) +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())
```

```
# 0.0
p5 <- ggplot(
  data = correlations2,
  mapping = aes(
    x = criterion,
    y = v11
  )
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE) +
  labs(
    title = "No Association",
    subtitle = expression(paste(italic("r"), " = ", ".0"))
  ) +
  theme_classic(
    base_size = 12) +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())

# 0.2
p6 <- ggplot(
  data = correlations2,
  mapping = aes(
    x = criterion,
    y = v13
  )
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE) +
  labs(
    title = "Weak Positive Association",
    subtitle = expression(paste(italic("r"), " = ", ".2"))
  ) +
  theme_classic(
    base_size = 12) +
```

```
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())

# 0.5
p7 <- ggplot(
  data = correlations2,
  mapping = aes(
    x = criterion,
    y = v16
  )
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE) +
  labs(
    title = "Moderate Positive Association",
    subtitle = expression(paste(italic("r"), " = ", ".5"))
  ) +
  theme_classic(
    base_size = 12) +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())

# 0.9
p8 <- ggplot(
  data = correlations2,
  mapping = aes(
    x = criterion,
    y = v20
  )
) +
  geom_point() +
  geom_smooth(
```

```
    method = "lm",
    se = FALSE) +
  labs(
    title = "Strong Positive Association",
    subtitle = expression(paste(italic("r"), " = ", ".9"))
  ) +
  theme_classic(
    base_size = 12) +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())

# 1.0
p9 <- ggplot(
  data = correlations2,
  mapping = aes(
    x = criterion,
    y = v21
  )
) +
  geom_point() +
  geom_smooth(
    method = "lm",
    se = FALSE) +
  labs(
    title = "Perfect Positive Association",
    subtitle = expression(paste(italic("r"), " = ", "1.0"))
  ) +
  theme_classic(
    base_size = 12) +
  theme(
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())

p1 + p2 + p3 + p4 + p5 + p6 + p7 + p8 + p9 +
  plot_layout(
```

```
    ncol = 3,
    heights = 1,
    widths = 1)
```

Figure 8.2 Interpretation of the Magnitude and Direction (Sign) of Correlation Coefficients.

Interactive visualizations by Kristoffer Magnusson on *p*-values and null-hypothesis significance testing are below:

- https://rpsychologist.com/correlation/ (archived at https://perma.cc/G8YR-VCM4)

8.4 Examples

8.4.1 Covariance

8.4.2 Pearson Correlation

8.4.3 Spearman Correlation

8.4.4 Nonlinear Correlation

8.4.5 Correlation Matrix

8.4.6 Correlogram

8.5 Correlation Does Not Imply Causation

As described in Section 6.3.2.1, correlation does not imply causation. There are several reasons (described in Section 6.3.2.1) that, just because `X` is correlated with `Y` does not necessarily mean that `X` causes `Y`. However, correlation can still be useful. In order for two processes to be causally related, they must be associated. That is, association is necessary but insufficient for causality.

8.6 Conclusion

Correlation is an index of the association between variables. The correlation coefficient (r) ranges from -1 to $+1$, and indicates the sign and magnitude of the association. Although correlation does not imply causation, identifying associations between variables can still be useful because association is a necessary (but insufficient) condition for causality.

8.7 Session Info

9

Multiple Regression

9.1 Getting Started

9.1.1 Load Packages

```
library("petersenlab")
library("tidyverse")
library("knitr")
```

9.2 Overview of Multiple Regression

Multiple regression examines the association between multiple predictor variables and one outcome variable. It allows obtaining a more accurate estimate of the unique contribution of a given predictor variable, by controlling for other variables (covariates).

Regression with one predictor variable takes the form of Equation 9.1:

$$y = \beta_0 + \beta_1 x_1 + \epsilon \tag{9.1}$$

where y is the outcome variable, β_0 is the intercept, β_1 is the slope, x_1 is the predictor variable, and ϵ is the error term.

A regression line is depicted in Figure 9.4.

Regression with multiple predictors—i.e., multiple regression—takes the form of Equation 9.2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon \tag{9.2}$$

where p is the number of predictor variables.

Figure 9.1 A Regression Best-Fit Line.

9.3 Components

- B = unstandardized coefficient: direction and magnitude of the estimate (original scale)
- β (beta) = standardized coefficient: direction and magnitude of the estimate (standard deviation scale)
- SE = standard error: uncertainty of unstandardized estimate

The unstandardized regression coefficient (B) is interpreted such that, for every unit change in the predictor variable, there is a ___ unit change in the outcome variable. For instance, when examining the association between age and fantasy points, if the unstandardized regression coefficient is 2.3, players score on average 2.3 more points for each additional year of age. (In reality, we might expect a nonlinear, inverted-U-shaped association between age and fantasy points such that players tend to reach their peak in the middle of their careers.) Unstandardized regression coefficients are tied to the metric of the raw data. Thus, a large unstandardized regression coefficient for two variables

may mean completely different things. Holding the strength of the association constant, you tend to see larger unstandardized regression coefficients for variables with smaller units and smaller unstandardized regression coefficients for variables with larger units.

Standardized regression coefficients can be obtained by standardizing the variables to *z*-scores so they all have a mean of zero and standard deviation of one. The standardized regression coefficient (β) is interpreted such that, for every standard deviation change in the predictor variable, there is a ___ standard deviation change in the outcome variable. For instance, when examining the association between age and fantasy points, if the standardized regression coefficient is 0.1, players score on average 0.1 standard deviation more points for each additional standard deviation of their year of age. Standardized regression coefficients—though not the case in all instances—tend to fall between [−1, 1]. Thus, standardized regression coefficients tend to be more comparable across variables and models compared to unstandardized regression coefficients. In this way, standardized regression coefficients provide a meaningful index of effect size.

9.4 Coefficient of Determination (R^2)

The coefficient of determination (R^2) reflects the proportion of variance in the outcome (dependent) variable that is explained by the model predictions: $R^2 = \frac{\text{variance explained in } Y}{\text{total variance in } Y}$. Various formulas for R^2 are in Equation 7.19. Larger R^2 values indicate greater accuracy. Multiple regression can be conceptualized with overlapping circles (similar to a venn diagram), where the non-overlapping portions of the circles reflect nonshared variance and the overlapping portions of the circles reflect shared variance, as in Figure 9.4.

One issue with R^2 is that it increases as the number of predictors increases, which can lead to overfitting if using R^2 as an index to compare models for purposes of selecting the "best-fitting" model. Consider the following example (adapted from Petersen (2024c)) in which you have one predictor variable and one outcome variable, as shown in Table 9.1.

Table 9.1 Example Data of Predictor (x1) and Outcome (y) Used for Regression Model.

y	x1
7	1
13	2
29	7

Table 9.1 Example Data of Predictor (x1) and Outcome (y) Used for Regression Model.

y	x1
10	2

Using the data, the best fitting regression model is: $y = 3.98 + 3.59 \cdot x_1$. In this example, the R^2 is 0.98. The equation is not a perfect prediction, but with a single predictor variable, it captures the majority of the variance in the outcome.

Now consider the following example where you add a second predictor variable to the data above, as shown in Table 9.2.

Table 9.2 Example Data of Predictors (x1 and x2) and Outcome (y) Used for Regression Model.

y	x1	x2
7	1	3
13	2	5
29	7	1
10	2	2

With the second predictor variable, the best fitting regression model is: $y = 0.00 + 4.00 \cdot x_1 + 1.00 \cdot x_2$. In this example, the R^2 is 1.00. The equation with the second predictor variable provides a perfect prediction of the outcome.

Providing perfect prediction with the right set of predictor variables is the dream of multiple regression. So, using multiple regression, we often add predictor variables to incrementally improve prediction. Knowing how much variance would be accounted for by random chance follows Equation 9.3:

$$E(R^2) = \frac{K}{n-1} \tag{9.3}$$

where $E(R^2)$ is the expected value of R^2 (the proportion of variance explained), K is the number of predictor variables, and n is the sample size. The formula demonstrates that the more predictor variables in the regression model, the more variance will be accounted for by chance. With many predictor variables and a small sample, you can account for a large share of the variance merely by chance.

As an example, consider that we have 13 predictor variables to predict fantasy performance for 43 players. Assume that, with 13 predictor variables, we

$$R^2 = \frac{\mathrm{D} + \mathrm{E} + \mathrm{G}}{\mathrm{A} + \mathrm{D} + \mathrm{E} + \mathrm{G}}$$

Figure 9.2 Conceptual Depiction of Proportion of Variance Explained (R^2) in an Outcome Variable (Y) by Multiple Predictors ($X1$ and $X2$) in Multiple Regression. The size of each circle represents the variable's variance. The proportion of variance in Y that is explained by the predictors is depicted by the areas in orange. The dark orange space (G) is where multiple predictors explain overlapping variance in the outcome. Overlapping variance that is explained in the outcome (G) will not be recovered in the regression coefficients when both predictors are included in the regression model. From Petersen (2024b) and Petersen (2024c).

explain 38% of the variance ($R^2 = .38; r = .62$). We explained a lot of the variance in the outcome, but it is important to consider how much variance could have been explained by random chance: $E(R^2) = \frac{K}{n-1} = \frac{13}{43-1} = .31$. We expect to explain 31% of the variance, by chance, in the outcome. So, 82% of the variance explained was likely spurious (i.e., $\frac{.31}{.38} = .82$). As the sample size increases, the spuriousness decreases.

To account for the number of predictor variables in the model, we can use a modified version of R^2 called adjusted R^2 (R^2_{adj}). Adjusted R^2 (R^2_{adj}) accounts for the number of predictor variables in the model, based on how much would be expected to be accounted for by chance to penalize overfitting. Adjusted R^2 (R^2_{adj}) reflects the proportion of variance in the outcome (dependent) variable that is explained by the model predictions over and above what would

be expected to be accounted for by chance, given the number of predictor variables in the model. The formula for adjusted R^2 (R^2_{adj}) is in Equation 9.4:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{9.4}$$

where p is the number of predictor variables in the model, and n is the sample size.

9.5 Overfitting

Statistical models applied to big data (e.g., data with many predictor variables) can *overfit* the data, which means that the statistical model accounts for error variance, which will not generalize to future samples. So, even though an overfitting statistical model appears to be accurate because it is accounting for more variance, it is not actually that accurate—it will predict new data less accurately than how accurately it accounts for the data with which the model was built. In the case of fantasy football analytics, this is especially relevant because there are hundreds if not thousands of variables we could consider for inclusion and many, many players when considering historical data.

Consider an example where you develop an algorithm to predict players' fantasy performance based on 2023 data using hundreds of predictor variables. To some extent, these predictor variables will likely account for true variance (i.e., signal) and error variance (i.e., noise). If we were to apply the same algorithm based on the 2023 prediction model to 2024 data, the prediction model would likely predict less accurately than with 2023 data. The regression coefficients in the

In Figure 9.3, the blue line represents the true distribution of the data, and the red line is an overfitting model:

```
set.seed(52242)

sampleSize <- 200
quadraticX <- rnorm(sampleSize)
quadraticY <- quadraticX ^ 2 + rnorm(sampleSize)
quadraticData <- cbind(quadraticX, quadraticY) %>%
  data.frame %>%
  arrange(quadraticX)

quadraticModel <- lm(
```

```
  quadraticY ~ quadraticX + I(quadraticX ^ 2),
  data = quadraticData)

quadraticNewData <- data.frame(
  quadraticX = seq(
    from = min(quadraticData$quadraticX),
    to = max(quadraticData$quadraticY),
    length.out = sampleSize))

quadraticNewData$quadraticY <- predict(
  quadraticModel,
  newdata = quadraticNewData)

loessFit <- loess(
  quadraticY ~ quadraticX,
  data = quadraticData,
  span = 0.01,
  degree = 1)

loessNewData <- data.frame(
  quadraticX = seq(
    from = min(quadraticData$quadraticX),
    to = max(quadraticData$quadraticY),
    length.out = sampleSize))

quadraticNewData$loessY <- predict(
  loessFit,
  newdata = quadraticNewData)

plot(
  x = quadraticData$quadraticX,
  y = quadraticData$quadraticY,
  xlab = "",
  ylab = "")

lines(
  quadraticNewData$quadraticY ~ quadraticNewData$quadraticX,
  lwd = 2,
  col = "blue")

lines(
  quadraticNewData$loessY ~ quadraticNewData$quadraticX,
  lwd = 2,
  col = "red")
```

Figure 9.3 Over-fitting Model in Red Relative to the True Distribution of the Data in Blue. From Petersen (2024b) and Petersen (2024c).

9.6 Covariates

Covariates are variables that you include in the statistical model to try to control for them so you can better isolate the unique contribution of the predictor variable(s) in relation to the outcome variable. Use of covariates examines the association between the predictor variable and the outcome variable when holding people's level constant on the covariates. Inclusion of confounds as covariates allows potentially gaining a more accurate estimate of the causal effect of the predictor variable on the outcome variable. Ideally, you want to include any and all confounds as covariates. As described in Section 6.3.2.1, confounds are third variables that influence both the predictor variable and the outcome variable and explain their association. Covariates are potentially (but not necessarily) confounds. For instance, you might include the player's age as a covariate in a model that examines whether a player's 40-yard dash time at the NFL Combine predicts their fantasy points in their rookie year, but it may not be a confound.

9.7 Multicollinearity

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. The problem of having multiple predictor variables that are highly correlated is that it makes it challenging to estimate the regression coefficients accurately.

Multicollinearity in multiple regression is depicted conceptually in Figure 9.4.

Figure 9.4 Conceptual Depiction of Multicollinearity in Multiple Regression. From Petersen (2024b) and Petersen (2024c).

Consider the following example adapted from Petersen (2024c) where you have two predictor variables and one outcome variable, as shown in Table 9.3.

Table 9.3 Example Data of Predictors (x1 and x2) and Outcome (y) Used for Regression Model.

y	x1	x2
9	2.0	4
11	3.0	6
17	4.0	8
3	1.0	2

Table 9.3 Example Data of Predictors (x1 and x2) and Outcome (y) Used for Regression Model.

y	x1	x2
21	5.0	10
13	3.5	7

The second predictor variable is not very good—it is exactly twice the value of the first predictor variable; thus, the two predictor variables are perfectly correlated (i.e., $r = 1.0$). This means that there are different prediction equation possibilities that are equally good—see Equations in Equation 9.5:

$$\begin{aligned} 2x_2 &= y \\ 0x_1 + 2x_2 &= y \\ 4x_1 &= y \\ 4x_1 + 0x_2 &= y \\ 2x_1 + 1x_2 &= y \\ 5x_1 - 0.5x_2 &= y \\ ... &= y \end{aligned} \tag{9.5}$$

Then, what are the regression coefficients? We do not know what are the correct regression coefficients because each of the possibilities fits the data equally well. Thus, when estimating the regression model, we could obtain arbitrary estimates of the regression coefficients with an enormous standard error around each estimate. In general, multicollinearity increases the uncertainty (i.e., standard errors and confidence intervals) around the parameter estimates. Any predictor variables that have a correlation above $\sim r = .30$ with each other could have an impact on the confidence interval of the regression coefficient. As the correlations among the predictor variables increase, the chance of getting an arbitrary answer increases, sometimes called "bouncing betas." So, it is important to examine a correlation matrix of the predictor variables before putting them in the same regression model. You can also examine indices such as variance inflation factor (VIF).

To address multicollinearity, you can drop a redundant predictor or you can also use principal component analysis or factor analysis of the predictors to reduce the predictors down to a smaller number of meaningful predictors. For a meaningful answer in a regression framework that is precise and confident, you need a low level of intercorrelation among predictors, unless you have a very large sample size.

10

Causal Inference

10.1 Getting Started

10.1.1 Load Packages

```
library("dagitty")
library("ggdag")
```

10.2 Correlation Does Not Imply Causation

As described in Section 6.3.2.1, there are several reasons why two variables, `X` and `Y`, might be correlated:

- `X` causes `Y`
- `Y` causes `X`
- `X` and `Y` are bidirectional: `X` causes `Y` and `Y` causes `X`
- a third variable (i.e., confound), `Z`, influences both `X` and `Y`
- the association between `X` and `Y` is spurious

10.3 Criteria for Causality

How do we know whether two processes are causally related? There are three criteria for establishing causality (Shadish et al., 2002):

1. The cause (e.g., the independent or predictor variable) temporally precedes the effect (i.e., the dependent or outcome variable).

2. The cause is related to (i.e., associated with) the effect.
3. There are no other alternative explanations for the effect apart from the cause.

The first criterion for establishing causality involves temporal precedence. In order for a cause to influence an effect, the cause must occur before the effect. For instance, if sports drink consumption influences player performance, the sports drink consumption (that is presumed to influence performance) must occur prior to the performance improvement. Establishing the first criterion eliminates the possibility that the association between the purported cause and effect reflects reverse causation. Reverse causation occurs when the purported effect is actually the cause of the purported cause, rather than the other way around. For instance, if sports drink consumption occurs only once, and it occurs only before and not after performance, then we have ruled out the possibility of reverse causation (i.e., that better performance causes players to consume sports drink).

The second criterion involves association. The purported cause must be associated with the purported effect. Nevertheless, as the maxim goes, "correlation does not imply causation." Just because two variables are correlated does not necessarily mean that they are causally related. However, correlation is useful because causality requires that the two processes be correlated. That is, correlation is a necessary but insufficient condition for causality. For instance, if sports drink consumption influences player performance, sports drink consumption must be associated with performance improvement.

The third criterion involves ruling out alternative reasons why the purported cause and effect may be related. As noted in Section 10.2, there are four reasons why `X` may be correlated with `Y`. If we meet the first criterion of causality, we have removed the possibility that `Y` causes `X` (i.e., reverse causality). To meet the third criterion of causality, we need to remove the possibility that the association reflects a third variable (confound) that influences both the cause and effect, and we need to remove the possibility that the association is spurious—the possibility that the association between the purported cause and effect is due to random chance.

There are multiple approaches to meeting the third criterion of causality, such as by use of experiments, longitudinal designs, control variables, within-subject designs, and genetically informed designs, as described in Section 10.4.

In general, to meet the third criterion of causality, one must consider the counterfactual. A *counterfactual* is what would have happened in the hypothetical scenario that the cause did not occur [i.e., what would have happened in the absence of the cause; Shadish et al. (2002)]. When engaging in causal inference, it is important to consider what would have happened if the hypothetical cause had actually not occurred. For instance, consider that we conduct an experiment to randomly assign some players to consume a sports

drink before a game and other players to drink only water. In this case, our treatment/intervention group is the group of players that consumed a sports drink. The control group is the group players that drank only water. Now, consider that the players in the treatment group outperform the players in the the control group in their football game. In such a study, we observe what *did happen* when players received a treatment. The counterfactual is knowledge of what *would have happened* to those same players if they simultaneously had not received treatment (Shadish et al., 2002). The true causal effect, then, is the difference between what did happen and what would have happened. However, we cannot observe a counterfactual. That is, we do not know for sure what would have happened to the players who received treatment if those same players had actually not received treatment. We have a control group, but the control group does not have the same players as the intervention group, and it is impossible for a person to simultaneously receive and not receive treatment.

So, our goal in working toward causal inference as scientists is to create reasonable approximations to this impossible counterfactual (Shadish et al., 2002). For instance, if using a between-subject design, we want the two groups to be equivalent in every possible way except whether or not they receive the treatment, so we might stratify each group to be equivalent in terms of age, weight, position, experience, skill, etc. Or, we might test the same people using an A-B-A-B within-subject design. In an A-B-A-B within-subject design, players receive no treatment at baseline (timepoint 1: game 1), receive the treatment at timepoint 2 (game 2), receive no treatment at timepoint 3 (game 3), and receive the treatment at timepoint 4 (game 4). Neither of these approximations is a true counterfactual. In the between-subject design, the players differ between the two groups, so we cannot know how the individuals who received the treatment would have performed if they had actually not received the treatment. In the A-B-A-B within-subject design, the players are the same, but they timepoints that they receive or do not receive the treatment differ, and there can be carryover effects from one condition to the next. For instance, consuming sports drinks before game 2 might also help them be better hydrated in general, including, for subsequent games. Thus, we cannot know how a player would have performed in game 1 with treatment or in game 2 without treatment, etc. Nevertheless, it is important to be aware of the counterfactual and to engage in counterfactual reasoning to consider what would have happened if the supposed cause had not occurred. Considering the counterfactual is important for designing closer approximations to the counterfactual in studies for stronger research designs and stronger causal inference.

10.4 Approaches for Causal Inference

10.4.1 Experimental Designs

As described in Section 6.3.1, experimental designs are designs in which participants are randomly assigned to one or more levels of the independent variable to observe its effects on the dependent variable. Experimental designs provide the strongest tests of causality because they can rule out reverse causation and third variables. For instance, by manipulating sports drink consumption before the player performs, they can eliminate the possibility that reverse causation explains the effect of the independent variable on the dependent variable. Second, through randomly assigning players to consume or not consume sports drink, this holds everything else constant (so long as the groups are evenly distributed according to other factors, such as their age, weight, etc.) and thus removes the possibility that third variable confounds explain the effect of the independent variable on the dependent variable.

10.4.2 Quasi-Experimental Designs

Although experimental designs provide the strongest tests of causality, manytimes they are impossible, unethical, or impractical to conduct. For instance, it would likely not be practical to randomly assign National Football League (NFL) players to either consume or not consume sports drink before their games. Players have their pregame rituals and routines and many would likely not agree to participate in such a study. Thus, we often rely on quasi-experimental designs such as natural experiments and observational/correlational designs.

We cannot directly test or establish causality from a non-experimental research design. Nevertheless, we can leverage various design features that, in combination with other studies using different research methods, collectively strengthen our ability to make causal inferences. For instance, there are are no experiments in humans showing that smoking causes cancer—randomly assigning people to smoke or not smoke would not be ethical. The causal inference that smoking causes cancer was derived from a combination of experimental studies in rodents and observational studies in humans.

10.4.2.1 Longitudinal Designs

Research designs can be compared in terms of their internal validity—the extent to which we can be confident about causal inferences. A cross-sectional association is depicted in Figure 10.1:

Figure 10.1 Cross-Sectional Association. T1 = Timepoint 1. From Petersen (2024b) and Petersen (2024c).

For instance, we might observe that sports drink consumptions is concurrently associated with better player performance. Among observational/correlational research designs, cross-sectional designs tend to have the weakest internal validity. For the reasons described in Section 10.2, if we observe a cross-sectional association between `X` (e.g., sports drink consumption) and `Y` (e.g., player performance), we have little confidence that `X` causes `Y`. As a result, longitudinal designs can be valuable for more closely approximating causality if an experimental designs is not possible. Consider a lagged association that might be observed in a longitudinal design, as in Figure 10.2, which is a slightly better approach than relying on cross-sectional associations:

For instance, we might observe that sports drink performance *before* the game is associated with better player performance *during* the game. A lagged association has somewhat better internal validity than a cross-sectional association because we have greater evidence of temporal precedence—that the influence of the predictor *precedes* the outcome because the predictor was assessed before the outcome and it shows a predictive association. However, part of the association between the predictor with later levels of the outcome could be due to prior levels of the outcome that are stable across time. That is, it could be that better player performance leads players to consume more sports drink and

Figure 10.2 Lagged Association. T1 = Timepoint 1. T2 = Timepoint 2. From Petersen (2024b) and Petersen (2024c).

that player performance is relatively stable across time. In such a case, it may be observed that sports drink consumption predicts later player performance even though player performance influences sports drink consumption, rather than the other way around Thus, consider an even stronger alternative—a lagged association that controls for prior levels of the outcome, as in Figure 10.3:

For instance, we might observe that sports drink performance *before* the game is associated with better player performance *during* the game, while controlling for prior player performance. A lagged association controlling for prior levels of the outcome has better internal validity than a lagged association that does not control for prior levels of the outcome. A lagged association that controls for prior levels further reduces the likelihood that the association owes to the reverse direction of effect, because earlier levels of the outcome are controlled. However, consider an even stronger alternative—lagged associations

Figure 10.3 Lagged Association, Controlling for Prior Levels of the Outcome. T1 = Timepoint 1. T2 = Timepoint 2. From Petersen (2024b) and Petersen (2024c).

that control for prior levels of the outcome and that simultaneously test each direction of effect, as depicted in Figure 10.4:

Lagged associations that control for prior levels of the outcome and that simultaneously test each direction of effect provide the strongest internal validity among observational/correlational designs. Such a design can help better clarify which among the variables is the chicken and the egg—which variable is more likely to be the cause and which is more likely to be the effect. If there are bidirectional effects, such a design can also help clarify the magnitude of each direction of effect. For instance, we can simultaneously evaluate the extent to which sports drink predicts later player performance (while controlling for prior performance) and the reverse—player performance predicting later sports drink consumption (while contorlling for prior sports drink consumption).

Figure 10.4 Lagged Association, Controlling for Prior Levels of the Outcome, Simultaneously Testing Both Directions Of Effect. T1 = Timepoint 1. T2 = Timepoint 2. From Petersen (2024b) and Petersen (2024c).

10.4.2.2 Within-Subject Analyses

Another design feature of longitudinal designs that can lead to greater internal validity is the use of within-subject analyses. Between-subject analyses, might examine, for instanc, whether players who consume more sports drink perform better on average compared to players who consume less sports drink. However, there are other between-person differences that could explain any observed between-subject associations between sports drink consumption and players performance. Another approach could be to apply within-subject analyses. For instance, you could examining whether, within the same individual, if a player consumes a sports drink, do they perform better compared to games in which they did not consume a sports drink. When we control for prior levels of the outcome in the prediction, we are evaluating whether the predictor is

associated with witin-person *change* in the outcome. Predicting within-person change provides stronger evidence consistent with causality because it uses the individual as their own control and controls for many time-invariant confounds (i.e., confounds that do not change across time). However, predicting within-person change does not, by itself, control for time-varying confounds. So, it can also be useful to control for time-varying confounds, such as by use of control variables.

10.4.2.3 Control Variables

One of the plausible alternatives to the inference that `X` causes `Y` is that there are third variable confounds that influence both `X` and `Y`, thus explaining why `X` and `Y` are associated, as depicted in Figures 6.3 and 10.10. Thus, another approach that can help increase internal validity is to include plausible confounds as control variables. For instance, if a third variable such as education level might be a confound that influences both sports drink consumption and player performance, you could include education level as a covariate in the model. Inclusion of a covariate attempts to control for the variable by examining the association between the predictor variable and the outcome variable while holding the covariate variables constant. For instance, such a model would examine whether, when accounting for education level, there is an association between sports drink consumption and player performance.

Failure to control for important third variables can lead to erroneous conclusions, as evidenced by the association depicted in Figure 10.5. In the example, if we did not control for gender, we would infer that there is a positive association between dosage and recovery probability. However, when we examine each men and women separately, we learn that the association between dosage and recovery probability is actually negative within each gender group. Thus, in this case, failure to control for gender would lead to false inferences about the association between dosage and recovery probability.

However, it can be problematic to control for variables indiscriminantly. The use of causal diagrams can inform which variables are important to be included as control variables, and—just as important—which variables not to include as control variables, as described in Section 10.5.

10.4.2.4 Genetically Informed Designs

Another approach to control for variables is to use genetically informed designs. Genetically informed designs allow controlling for potential genetic effects in order to more closely approximate the contributions of various environmental effects. Genetically informed designs exploit differing degrees of genetic relatedness among participants to capture the extent to which genetic factors may contribute to an outcome. The average per-

Figure 10.5 Example Where Failing to Control for a Variable (In This Case, Gender) Would Lead to False Inferences. In this example, the association between dosage and recovery probability is positive at the population level, but the association is negative among men and women separately. (Figure reprinted from Kievit et al. (2013), Figure 1, p. 2. Kievit, R., Frankenhuis, W., Waldorp, L., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology*, *4*(513). https://doi.org/10.3389/fpsyg.2013.00513)

cent of DNA shared between people of varying relationships is provided in Table 10.1 (https://isogg.org/wiki/Autosomal_DNA_statistics; archived at https://perma.cc/MK3D-DST8):

Table 10.1 Average Percent of Autosomal DNA Shared by Pairs of Relatives by Relationship Type.

Relationship	Average Percent of Autosomal DNA Shared by Pairs of Relatives
Monozygotic ("identical") twins	100%
Dizygotic ("fraternal") twins	50%
Parent/child	50%
Full siblings	50%
Grandparent/grandchild	25%
Aunt-or-uncle/niece-or-nephew	25%
Half-siblings	25%
First cousin	12.5%
Great-grandparent/great-grandchild	12.5%

For instance, researchers may compare monozygotic twins versus dizygotic twins in some outcome—a so-called "twin study". It is assumed that the trait/outcome is attributable to genetic factors to the extent that the monozygotic twins (who share 100% of their DNA) are more similar in the trait or outcome compared to the dizygotic twins (who share on average 50% of their DNA). Alternatively, researchers could compare full siblings versus half-siblings, or they could compare full siblings versus first cousins.

Genetically informed designs are not as relevant for fantasy football analytics, but they are useful to present as one of various design features that researchers can draw upon to strengthen their ability to make causal inferences.

10.5 Causal Diagrams

10.5.1 Overview

A key tool when describing a research question or hypothesis is to create a conceptual depiction of the hypothesized causal processes. A causal diagram

depicts the hypothesized causal processes that link two or more variables. A common form of causal diagrams is the directed acyclic graph (DAG). DAGs provide a helpful tool to communicate about causal questions and help identify how to avoid bias (i.e., over-estimation) in associations between variables due to confounding (i.e., common causes) (Digitale et al., 2022). For instance, from a DAG, it is possible to determine what variables it is important to control for in order to get unbiased estimates of the association between two variables of interest. To create DAGs, you can use the `R` package `dagitty` (Textor et al., 2017) or the associated browser-based extension, DAGitty: https://dagitty.net (archived at https://perma.cc/U9BY-VZE2). Examples of various causal diagrams that could explain why `X` is associated with `Y` are in Figures 10.6, 10.8 and 10.10.

```
XCausesY <- dagitty::dagitty("dag{
  X -> Y
}")

plot(dagitty::graphLayout(XCausesY))

dagitty::impliedConditionalIndependencies(XCausesY)
```

Figure 10.6 Causal Diagram (Directed Acyclic Graph) Depicting `X` Causing `Y`.

Here is an alternative way of specifying the same diagram (more similar to `lavaan` syntax):

```
XCausesY_alt <- ggdag::dagify(
  Y ~ X
)

#plot(XCausesY_alt) # this creates the same plot as above
ggdag::ggdag(XCausesY_alt) + theme_dag_blank()
```

Figure 10.7 Causal Diagram (Directed Acyclic Graph) Depicting `X` Causing `Y`.

```
YCausesX <- dagitty::dagitty("dag{
  Y -> X
}")

plot(dagitty::graphLayout(YCausesX))

dagitty::impliedConditionalIndependencies(YCausesX)
```

Figure 10.8 Causal Diagram (Directed Acyclic Graph) Depicting Reverse Causation: `Y` Causing `X`.

Here is an alternative way of specifying the same diagram (more similar to `lavaan` syntax):

```
YCausesX_alt <- ggdag::dagify(
  X ~ Y
)

#plot(YCausesX_alt) # this creates the same plot as above
ggdag::ggdag(YCausesX_alt) + theme_dag_blank()
```

Figure 10.9 Causal Diagram (Directed Acyclic Graph) Depicting Reverse Causation: `Y` Causing `X`.

```
ZCausesXandY <- dagitty::dagitty("dag{
  Z -> Y
  Z -> X
  X <-> Y
}")

plot(dagitty::graphLayout(ZCausesXandY))
```

Figure 10.10 Causal Diagram (Directed Acyclic Graph) Depicting a Third Variable Confound, `Z`, Causing `X` and `Y`, Thus Explaining Why `X` and `Y` are associated.

Here is an alternative way of specifying the same diagram (more similar to `lavaan` syntax):

```
ZCausesXandY_alt <- ggdag::dagify(
  X ~ Z,
  Y ~ Z,
  X ~~ Y
)

#plot(ZCausesXandY_alt) # this creates the same plot as above
ggdag::ggdag(ZCausesXandY_alt) + theme_dag_blank()
```

Figure 10.11 Causal Diagram (Directed Acyclic Graph) Depicting a Third Variable Confound, `Z`, Causing `X` and `Y`, Thus Explaining Why `X` and `Y` are associated.

Consider another example in Figure 10.12:

```
mediationDag <- dagitty::dagitty("dag{
  X -> M1
  X -> M2
  M1 -> Y
  M2 -> Y
  M1 <-> M2
}")

plot(dagitty::graphLayout(mediationDag))
```

Figure 10.12 Causal Diagram (Directed Acyclic Graph).

```
dagitty::impliedConditionalIndependencies(mediationDag)
```

```
X _||_ Y | M1, M2
```

```
dagitty::adjustmentSets(
  mediationDag,
  exposure = "M1",
  outcome = "Y",
  effect = "total")
```

```
{ M2 }
```

In this example, `X` influences `Y` via `M1` and `M2` (i.e., multiple mediators), and `M1` is also associated with `M2`. The `dagitty::impliedConditionalIndependencies()` function identifies variables in the causal diagram that are conditionally independent (i.e., uncorrelated) after controlling for other variables in the model. For this causal diagram, `X` is conditionally independent with `Y` because `X` is independent with `Y` when controlling for `M1` and `M2`.

The `dagitty::adjustmentSets()` function identifies variables that would be necessary to control for (i.e., to include as covariates) in order to identify an unbiased estimate of the association (whether the total effect, i.e., `effect = "total"`; or the direct effect, i.e., `effect = "direct"`) between two variables

(`exposure` and `outcome`). In this case, to identify the unbiased association between `M1` and `Y`, it is important to control for `M2`.

Here is an alternative way of specifying the same diagram (more similar to `lavaan` syntax):

```
mediationDag_alt <- ggdag::dagify(
  M1 ~ X,
  M2 ~ X,
  Y ~ M1,
  Y ~ M2,
  M1 ~~ M2
)

#plot(mediationDag_alt) # this creates the same plot as above
ggdag::ggdag(mediationDag_alt) + theme_dag_blank()
```

Figure 10.13 Causal Diagram (Directed Acyclic Graph).

10.5.2 Confounding

Confounding occurs when two variables—that are both caused by another variable(s)—have a spurious or noncausal association (D'Onofrio et al., 2020). That is, two variables share a common cause, and the common cause leads the variables to be associated even though they are not causally related. The

common cause—i.e., the variable that influences the two variables—is known as a confound (or confounder). An example of confounding is depicted in Figure 10.14:

```
confounding <- ggdag::confounder_triangle(
  x = "Player Endurance",
  y = "Field Goals Made",
  z = "Stadium Altitude")

confounding %>%
  ggdag(
    text = FALSE,
    use_labels = "label") +
  theme_dag_blank()
```

Figure 10.14 Causal Diagram (Directed Acyclic Graph) Example of Confounding.

```
dagitty::impliedConditionalIndependencies(confounding)
```

```
x _||_ y | z
```

The output indicates that player endurance (`X`) and field goals made (`Y`) are conditionally independent when accounting for stadium altitude (`Z`). *Conditional independence* refers to two variables being unassociated when controlling for other variables.

```
dagitty::adjustmentSets(
  confounding,
  exposure = "x",
  outcome = "y",
  effect = "total")
```

```
{ z }
```

The output indicates that, to obtain an unbiased estimate of the causal association between two variables, it is necessary to control for any confounding (Lederer et al., 2019). That is, to obtain an unbiased estimate of the causal association between player endurance (`X`) and field goals made (`Y`), it is necessary to control for stadium altitude (`Z`).

10.5.3 Mediation

Mediation can be divided into two types: full and partial. In full mediation, the mediator(s) fully account for the effect of the predictor variable on the outcome variable. In partial mediation, the mediator(s) partially but do not fully account for the effect of the predictor variable on the outcome variable.

10.5.3.1 Full Mediation

An example of full mediation is depicted in Figure 10.15:

```
full_mediation <- ggdag::mediation_triangle(
  x = "Coaching Quality",
  y = "Fantasy Points",
  m = "Player Preparation")

full_mediation %>%
  ggdag(
    text = FALSE,
    use_labels = "label") +
  theme_dag_blank()
```

Figure 10.15 Causal Diagram (Directed Acyclic Graph) Example of Full Mediation.

```
dagitty::impliedConditionalIndependencies(full_mediation)
```

```
x _||_ y | m
```

In full mediation, `X` and `Y` are conditionally independent when accounting for the mediator (`Z`). In this case, coaching quality (`X`) and fantasy points (`Y`) are conditionally independent when accounting for player preparation (`M`). In other words, in this example, player preparation is the mechanism that fully (i.e., 100%) accounts for the effect of coaching quality on players' fantasy points.

```
dagitty::adjustmentSets(
  full_mediation,
  exposure = "x",
  outcome = "y",
  effect = "direct")
```

```
{ m }
```

The output indicates that, to obtain an unbiased estimate of the *direct* causal association between coaching quality (`X`) and fantasy points (`Y`) (i.e., the effect that is *not* mediated through intermediate processes), it is necessary to control for player preparation (`M`).

```
dagitty::adjustmentSets(
  full_mediation,
  exposure = "x",
  outcome = "y",
  effect = "total")
```

```
{}
```

However, to obtain an unbiased estimate of the *total* causal association between coaching quality (`X`) and fantasy points (`Y`) (i.e., including the portion of the effect that is mediated through intermediate processes), it is important *not* to control for player preparation (`M`). When the goal is to understand the (total) causal effect of coaching quality (`X`) and fantasy points (`Y`), controlling for the mediator (player preparation; `M`) would be inappropriate because doing so would remove the causal effect, thus artificially reducing the estimate of the association between coaching quality (`X`) and fantasy points (`Y`) (Lederer et al., 2019).

10.5.3.2 Partial Mediation

An example of partial mediation is depicted in Figure 10.16:

```
partial_mediation <- ggdag::mediation_triangle(
  x = "Coaching Quality",
  y = "Fantasy Points",
  m = "Player Preparation",
  x_y_associated = TRUE)

partial_mediation %>%
  ggdag(
    text = FALSE,
    use_labels = "label") +
  theme_dag_blank()
```

Figure 10.16 Causal Diagram (Directed Acyclic Graph) Example of Partial Mediation.

```
dagitty::impliedConditionalIndependencies(partial_mediation)
```

In partial mediation, `X` and `Y` are *not* conditionally independent when accounting for the mediator (`Z`). In this case, coaching quality (`X`) and fantasy points (`Y`) are still associated when accounting for player preparation (`M`). In other words, in this example, player preparation is a mechanism that partially but does not fully account for the effect of coaching quality on players' fantasy points. That is, there are likely other mechanisms, in addition to player preparation, that collectively account for the effect of coaching quality on fantasy points. For instance, coaching quality could also influence player fantasy points through better play-calling.

```
dagitty::adjustmentSets(
  partial_mediation,
  exposure = "x",
  outcome = "y",
  effect = "direct")
```

```
{ m }
```

As with full mediation, the output indicates that, to obtain an unbiased estimate of the *direct* causal association between coaching quality (`X`) and fantasy

points (`Y`) (i.e., the effect that is *not* mediated through intermediate processes), it is necessary to control for player preparation (`M`).

```
dagitty::adjustmentSets(
  partial_mediation,
  exposure = "x",
  outcome = "y",
  effect = "total")
```

```
{}
```

However, as with full mediation, to obtain an unbiased estimate of the *total* causal association between coaching quality (`X`) and fantasy points (`Y`) (i.e., including the portion of the effect that is mediated through intermediate processes), it is important *not* to control for player preparation (`M`). When the goal is to understand the (total) causal effect of coaching quality (`X`) and fantasy points (`Y`), controlling for a mediator (player preparation; `M`) would be inappropriate because doing so would remove the causal effect, thus artificially reducing the estimate of the association between coaching quality (`X`) and fantasy points (`Y`) (Lederer et al., 2019).

10.5.4 Collider Bias

Collision occurs when two variables influence a third variable, the collider (D'Onofrio et al., 2020). That is, a collider is a variable that is caused by two other variables (i.e., a common effect). An example collision is depicted in Figures 10.17 and 10.18:

```
colliderBias1 <- ggdag::collider_triangle(
  x = "Diet",
  y = "Coaching Strategy",
  m = "Injury Status")

colliderBias1 %>%
  ggdag(
    text = FALSE,
    use_labels = "label") +
  theme_dag_blank()
```

Figure 10.17 Causal Diagram (Directed Acyclic Graph) Example of a Collision with a Collider (Injury Status).

```
dagitty::impliedConditionalIndependencies(colliderBias1)
```

```
x _||_ y
```

In this example collision, diet (`X`) and coaching strategy (`Y`) are independent.

```
dagitty::adjustmentSets(
  colliderBias1,
  exposure = "x",
  outcome = "y",
  effect = "total")
```

```
 {}
```

As the output indicates, we should not control for the collider when examining the association between the two causes of the collider. That is, we should not control for injury status (`M`) when examining the association between diet (`X`) and coaching strategy. Controlling for the collider leads to confounding and can artificially induce an association between the two causes of the collider despite no causal association between them (Lederer et al., 2019). Obtaining a distorted or artificial association between two variables due to inappropriately controlling for a collider is known as *collider bias*.

Consider another example:

```
colliderBias2 <- ggdag::collider_triangle(
  x = "Coaching Quality",
  y = "Player Preparation",
  m = "Fantasy Points",
  x_y_associated = TRUE)

colliderBias2 %>%
  ggdag(
    text = FALSE,
    use_labels = "label") +
  theme_dag_blank()
```

Figure 10.18 Causal Diagram (Directed Acyclic Graph) Example of Collider Bias.

```
dagitty::impliedConditionalIndependencies(colliderBias2)
```

In this example of collider bias, there are no conditional independencies.

```
dagitty::adjustmentSets(
  colliderBias2,
  exposure = "x",
  outcome = "y",
  effect = "total")
```

```
{}
```

Again, it would be important not to control for the collider, fantasy points (`M`), when examining the association between coaching quality (`X`) and player preparation (`Y`). In this case, controlling for the collider would remove some of the causal effect of coaching quality on player preparation and could lead to an artificially smaller estimate of the causal effect between coaching quality and player preparation.

10.5.4.1 M-Bias

Collider bias may also occur when neither variable of interest is a direct cause of the collider (Lederer et al., 2019). M-bias is a form of collider bias that occurs when two variables that are not causally related, `A` and `B`, both influence a collider, `M`, and each (`A` and `B`) also influences a separate variable—e.g., `A` influences `X` and `B` influences `Y`. M-bias is so-named from the M-shape of the DAG. An example of M-bias is depicted in Figure 10.19:

```
mBias <- ggdag::m_bias(
  x = "Number of Media Articles About the Team",
  y = "Fantasy Points",
  a = "Team Record",
  b = "Coaching Quality",
  m = "Fan Attendance at Game")

mBias %>%
  ggdag(
    text = FALSE,
    use_labels = "label") +
  theme_dag_blank()
```

Figure 10.19 Causal Diagram (Directed Acyclic Graph) Example of M-Bias.

In this example, fan attendance is the collider that is influenced separately by the team record and the coaching quality. This is a fictitious example for purposes of demonstration; in reality, coaching quality influences the team's record.

```
dagitty::impliedConditionalIndependencies(mBias)
```

```
a _||_ b
a _||_ y
b _||_ x
m _||_ x | a
m _||_ y | b
x _||_ y
```

As the output indicates, there are several conditional independencies.

```
dagitty::adjustmentSets(
  mBias,
  exposure = "x",
  outcome = "y",
  effect = "total")
```

```
{}
```

It is important not to control for the collider (fan attendance). If you control for the collider, you can induce an artificial association between team record and coaching quality. Moreover, because doing so induces an artificial association between team record and coaching quality, it can also induce an artificial association between the effects of team record and coaching quality: number of media articles about the team and fantasy points, respectively. That is, controlling for the collider can lead to an artificial association between `X` and `Y` that does not reflect a causal process.

10.5.4.2 Butterfly Bias

Butterfly bias occurs when both confounding and M-bias are present. Butterfly bias (aka bow-tie bias) is so-named from the butterfly shape of the DAG. In butterfly bias, the following criteria are met:

- Two variables (`A` and `B`) influence a collider (`M`).
- The collider influences two variables, `X` and `Y`.
- `A` also influences `X`.
- `B` also influences `Y`.
- `A` and `B` are not causally related.
- `X` and `Y` are not causally related.

Or, more succinctly:

- `A` influences `M` and `X`.
- `B` influences `M` and `Y`.
- `M` influences `X` and `Y`.

In butterfly bias, the collider (`M`) is also a confound. That is, a variable is both influenced by two variables and influences two variables. An example of butterfly bias is depicted in Figure 10.20:

```
butterflyBias <- ggdag::butterfly_bias(
  x = "Off-field Behavior",
  y = "Fantasy Points",
  a = "Diet",
  b = "Coaching Quality",
  m = "Mental Health")

butterflyBias %>%
  ggdag(
    text = FALSE,
    use_labels = "label") +
  theme_dag_blank()
```

Figure 10.20 Causal Diagram (Directed Acyclic Graph) Example of Butterfly Bias.

In this case, players' mental health is a collider of their diet and the quality of the coaching they receive. In addition, players' mental health is a confound of their off-field behavior and fantasy points.

```
dagitty::impliedConditionalIndependencies(butterflyBias)
```

```
a _||_ b
a _||_ y | b, m
b _||_ x | a, m
x _||_ y | b, m
x _||_ y | a, m
```

As the output indicates, there are several conditional independencies.

```
dagitty::adjustmentSets(
  butterflyBias,
  exposure = "x",
  outcome = "y",
  effect = "total")
```

```
{ b, m }
{ a, m }
```

When dealing with a collider that is also a confound, controlling for either set, `B` and `M` or `A` and `M`, will provide an unbiased estimate of the association between `X` and `Y`. In this case, controlling for either a) coaching quality and mental health or b) diet and mental health—but not both sets—will yield an unbiased estimate of the association between off-field behavior and fantasy points.

10.5.5 Selection Bias

Selection bias occurs when the selection of participants or their inclusion in analyses depends on the variables being studied. For instance, if you are conducting a study on the extent to which sports drink consumption influences fantasy points, there would be selection bias if players are less likely to participate in the study if they score fewer fantasy points.

Now, consider a study in which you conduct a randomized controlled trial (RCT; i.e., an experiment) to evaluate the effect of a new medication on player performance. You randomly assign some players to take the medication and other players to take a placebo. Assume the new medication has side effects and leads many of the players who take it to drop out of the study. This is an example of attrition bias (i.e., systematic attrition). If you were to exclude these individuals from your analysis, it may make it appear that the medication led to better performance, because the players who experienced the side effect (and worse performance) dropped out of the study. Hence, conducting an analysis that excludes these players from the analysis would involve selection bias.

10.6 Conclusion

There are three criteria for establishing causality: 1) the cause precedes the effect. 2) The cause is related to the effect. 3) There are no other alternative explanations for the effect apart from the cause. In general, it is important to be aware of the counterfactual and to consider what would have happened if the supposed cause had not occurred. Various experimental and quasi-experimental designs and approaches can be leveraged to more closely approximate causal inferences. Longitudinal designs, within-subject analyses, inclusion of control variables, and genetically informed designs are all quasi-experimental designs that afford the researcher greater control over some possible third variable confounds. Causal diagrams can be a useful tool for identifying the proper variables to control for (and those not to control for). When confounding exists, it is important to control for the confound(s). It is

important not to control for mediators when interested in the total effect of the predictor variable on the outcome variable. When there is a collision, it is important not to control for the collider unless the collider is also a confound.

11

Heuristics and Cognitive Biases in Prediction

11.1 Getting Started

11.1.1 Load Packages

```
library("tidyverse")
```

11.2 Overview

When considering judgment and prediction, it is important to consider psychological concepts, including heuristics and cognitive biases. In the modern world of big data, research and society need people who know how to make sense of the information around us. Given humans' cognitive biases, it is valuable to leverage more objective approaches than relying on our "gut" and intutition. Statistical approaches can be a more objective way to identify systematic patterns.

Statistical analysis—and science more generally—is a process to the pursuit of knowledge. An *epistemology* is an approach to knowledge. Science is perhaps the best approach (epistemology) that society has to approximate truth. Unlike other approaches to knowledge, science relies on empirical evidence and does not give undue weight to anecdotal evidence, intuition, tradition, or authority.

Per Petersen (2024c), here are the characteristics of science that distinguish it from pseudoscience:

1. Risky hypotheses are posed that are falsifiable. The hypotheses can be shown to be wrong.
2. Findings can be replicated independently by different research groups and different methods. Evidence converges across studies and methods.
3. Potential alternative explanations for findings are specified and examined empirically (with data).
4. Steps are taken to guard against the undue influence of personal beliefs and biases.
5. The strength of claims reflects the strength of evidence. Findings and the ability to make judgments or predictions are not overstated. For instance, it is important to present the degree of uncertainty from assessments with error bars or confidence intervals.
6. Scientifically supported measurement strategies are used based on their psychometrics, including reliability and validity.

Nevertheless, statistical analysis is not purely objective and is not a panacea. Science is a human enterprise—it is performed by humans each of whom has their own biases. For instance, cognitive biases can influence how people interpret statistics. As a result, the findings from any given study may be incorrect. Thus, it would be imprudent to make decisions based solely on the results of one study. That is why we wait for findings to be independently replicated by different groups of researchers using different methods.

If a research team publishes flashy new and exciting findings, other researchers have an incentive to disprove the prior findings. Thus, we have more confidence if findings stand up to scrutiny from independent groups of researchers. We also draw upon meta-analyses—studies of many studies, to summarize the results of many studies and not just the findings from any single study that may not replicate. In this way, we can identify which findings are robust and most likely true versus the findings that fail to replicate. Thus, despite its flaws like any other human enterprise, science is a self-correcting process in which the long arc bends toward truth.

In our everyday lives, humans are presented with overwhelming amounts of information. Because human minds cannot parse every piece of information equally, we tend to take mental shortcuts, called *heuristics*. These mental shortcuts can be helpful. They reduce our mental load and can help us make quick judgments to stay alive or to make complex decisions in the face of uncertainty. However, these mental shortcuts can also lead us astray and to make systematic errors in our judgments and predictions. *Cognitive biases* are

systematic errors in thinking. *Fallacies* are forms of flawed reasoning.

11.3 Examples of Heuristics

As described above, heuristics are mental shortcuts that people use to handle the overwhelming amount of information to process. Three important heuristics used in judgment and prediction in the face of uncertainty include (Tversky & Kahneman, 1974):

- availability heuristic
- representativeness heuristic
- anchoring and adjustment heuristic

11.3.1 Availability Heuristic

The *availability heuristic* refers to the tendency for a person's judgments or predictions about the frequency or probability of something to be made based on how readily instances can be brought to mind. For instance, when making fantasy predictions about a player, more recent big performance games may more easily come to mind compared to lower-scoring games and games that occurred longer ago. Thus, a manager may be more inclined to pick players to start who had more recent, stronger performances rather than players who have higher long-term averages.

11.3.2 Representativeness Heuristic

The *representativess heuristic* refers to the tendency for a person's judgments or predictions about individuals to be made based on how similar the individual is to the person's existing mental prototypes. For instance, when coming out of college, Tight End Kyle Pitts drew comparisons[1] to the "LeBron James" of Tight Ends (archived at https://perma.cc/JQB5-XPVL). The idea that his athletic profile leads him to be similar to the prototype of LeBron James may have led him to be too highly drafted by fantasy managers in his first seasons.

The representativeness heuristic has been observed in gambling markets for predicting team wins in the National Football League (NFL) (Woodland & Woodland, 2015) and in decision making in fantasy soccer (Kotrba, 2020).

[1] https://247sports.com/article/kyle-pitts-lebron-james-2021-nfl-draft-florida-gators-football-163882176

11.3.3 Anchoring and Adjustment Heuristic

The *anchoring and adjustment heuristic* refers to the tendency for a person's judgments or predictions to be made with a reference point—an anchor—as a starting point from which they adjust their estimates upward or downward. The anchor is often inaccurate and given too much weight in the person's calculation, and too little adjustment is made to the anchor. For instance, a manager is trying to predict how many fantasy points a top Running Back may score. The player scored 300 fantasy points last season, but the team added a stronger backup Running Back and changed the Offensive Coordinator to be a more pass-heavy offense. The manager may use 300 fantasy points as an anchor (based on the player's performance last season), and may adjust downward 15 points to account for the offseason changes. However, it is possible that this downward adjustment is insufficient to account not only for the offseasons changes but also for potential regression effects. Regression effects are discussed further in Section 11.5.2.

11.4 Examples of Cognitive Biases

As described above, cognitive biases are systematic errors in thinking. Cognitive biases are often due to the use of heuristics. Examples of cognitive biases that result from one or more of these heuristics include:

- overconfidence bias
- confirmation bias
- recency bias
- hindsight bias
- loss aversion bias
- endowment bias
- bandwagon effect bias
- Dunning–Kruger effect bias

11.4.1 Overconfidence Bias

In general, people tend to be overconfident in their judgments and predictions. *Overconfidence bias* is the tendency for a person to have greater confidence in their abilities (including judgments and predictions) than is objectively warranted. There are three general ways that overconfidence has been identified (Moore & Healy, 2008):

1. *overestimation* of one's actual performance
2. *overplacement* of one's performance relative to others
3. *overprecision* in one's beliefs/jugments/predictions

Overestimation involves believing that one will perform better than one actually performs. Overestimation can be identified with a calibration plot of the predicted performance versus actual performance, where the person's predicted performance is systematically higher (in at least some cases) than their actual performance. Overestimation corresponds to the "overprediction" form of miscalibration.

Overplacement involves believing that one is better than others or will perform better than others, even when they do not. For instance, it is a common finding that more than half of people believe they are "above average" (i.e., above the median), even though that is statistically impossible. This calls to mind the fictitious Lake Wobegon in the radio show *A Prararie Home Companion*, "where all the women are strong, all the men are good-looking, and all the children are above average."

Overprecision involves expressing excessive certainty regarding the accuracy of one's beliefs/judgments/predictions. For instance, if when a given meteorologist says it will rain 80% of the time, it actually rains 30% of the time, the meteorologist's predictions are overprecise. Likewise, if the weather forecast says it will rain 10% of the time and it actually rains 30% of the time, the predictions are also overprecise because the forecaster is expressing stronger confidence than is warranted that it will not rain. Overprecision can be identified with a calibration plot of the predicted probabilities versus the actual probabilities. Overprecision corresponds to the "overextremity" form of miscalibration.

Overestimation and overprecision are studied in various ways. Typically, people are asked about (a) whether an event will occur or (b) the likelihood that the event will occur, across many events. For the former (approach "a"), people may be asked to make a dichotomous judgment or prediction, by responding to the question: e.g., "Will it rain tomorrow? [YES/NO]". They will then rate their confidence (as a percentage) in their answer (0–100%). They would make each of these two ratings for each of many events. Then, we can evaluate, for a given respondent, the degree to which the probabilistic estimate of an event reflects the true underlying probability of the event. For instance, for a given respondent (and for respondents in general), for the events when the respondent says they are 80% confident an event (e.g., rain) will occur, does the event actually occur around 80% of the time? For the latter (approach "b"), people may indicate the likelihood that the event will occur, by responding to the question: "How likely is it that it will rain tomorrow? (0–100%)". Then, we can evaluate, for instance, for the events when the respondent says that an event (e.g., rain) is 80% likely to occur, does the event actually occur around 80% of the time?

A fantasy manager may be even more likely to exhibit overconfidence if they previously performed well or won their league, for which luck and random chance plays an important role. Indeed, it is estimated that nearly half (~45%) of the variability in fantasy football performance is estimated to be luck [and around 55% due to skill; Getty et al. (2018)]. A manager who won their league in the prior season may believe they will perform better than they actually will (overestimation), will perform better than average (overplacement), and may hold excessive confidence regarding the accuracy of their predictions about which players will perform well or poorly (overprecision). These various types of overconfidence may lead them to draft high-risk players based on gut feeling, neglecting statistical analysis and expert consensus.

Players' performance in fantasy football, and human behavior more generally, is complex and multiply determined (i.e., is influenced by many factors). Despite the bluster of so-called experts who pretend to know more than they can know, no one can consistently and accurately predict how all players will perform. Remain humble in your predictions; do not be more confident than is warranted. If you approach the task of prediction with humility, you may be more able to be flexible and more willing to consider other players who you can draft for good value.

11.4.2 Confirmation Bias

Confirmation bias is the tendency for people to search for, interpret, and remember information that confirms one's beliefs, as opposed to information that might disconfirm one's beliefs. The result of confirmation bias is that people are unlikely to change their minds about something that they have a pre-existing belief about, because they tend to look only for information that supports their pre-existing beliefs. For instance, if you believe that a particular player is a strong breakout candidate to be a sleeper, you may be more likely to pay attention to evidence that supports that the player will breakout and may be less likely to pay attention to evidence that indicates the player may struggle.

As a budding empiricist, you should actively seek out information that challenges or disconfirms your beliefs and work to incorporate it into your beliefs. Do your best to go into observation, data analysis, and data interpretation with an open mind.

11.4.3 Recency Bias

Recency bias is the tendency to weigh recent events more than earlier ones. For instance, a manager might observe that a Running Back on the waiver wire performed well in the last two games. *Recency bias* may lead the manager

to pick up the player, overvaluing their recent performance. For instance, the manager may not have adequately weighed the player's overall season performance and the fact that the starting Running Back is returning to the lineup from injury, and that is why the player received more carries in the past two games (i.e., in place of the injured starter).

11.4.4 Hindsight Bias

"Hindsight is 20/20." – Idiom

Hindsight bias is the tendency to perceive that past events were more predictable than they were. People tend to remember the succeess of their predictions and forget the failures of their predictions. For instance, if a third-string Quarterback has a breakout game, a fantasy manager may claim that they "knew it all along" that the player was going to breakout, despite not having picked up the player. That same manager may forget the many other predictions they had that did not come true.

11.4.5 Loss Aversion Bias

Loss aversion bias is the tendency to avoid losses rather than acquiring equivalent gains. Loss aversion is exemplified when teams play conservatively so as "not to lose" instead of "to win." In fantasy football, loss aversion may lead managers to start or hold onto underperforming high drafts for too long instead of a starting a more promising player out of fear of losing potential value from their initial investment.

11.4.6 Endowment Bias

Endowment bias is the tendency to overvalue merely because one owns it. For instance, a manager might overvalue a player they drafted in the first round, refusing to trade them even if they could get a better-performing player in return.

11.4.7 Bandwagon Effect Bias

The *bandwagon effect bias* is the tendency to do or believe things because other people are. It involves social conformity. For instance, consider if a rookie Wide Receiver has a breakout game and he is picked up in many fantasy leagues. A given manager might pick up the player because the player is frequently being picked up in many fantasy leagues, without evaluating whether the player's success is sustainable.

11.4.8 Dunning–Kruger Effect Bias

> "The more you know, the more you know you don't know." – Anonymous

The *Dunning–Kruger effect bias* is the tendency for people with low ability/competency in a task to overestimate their ability. The Dunning–Kruger effect is depicted in Figures 11.1 and 11.2. For instance, consider a new fantasy manager who experiences some initial wins (often called "beginner's luck"). They may attribute their successes to their skill rather than to luck. Their overconfidence may lead them to believe they can win the league without much preparation.

11.5 Examples of Fallacies

As described above, fallacies are mistaken beliefs and flawed reasoning. Fallacies are often due to the use of heuristics and to cognitive biases. Examples of fallacies include:

- base rate fallacy (aka base rate neglect)
- regression fallacy
- sunk cost fallacy
- hot hand fallacy
- gambler's fallacy
- conditional probability fallacy

Figure 11.1 Dunning–Krueger Effect: Confidence as a Function of Competence. Adapted from https://commons.wikimedia.org/wiki/File:Effet_Dunning-Kruger.svg.

Figure 11.2 Dunning–Krueger Effect: Perceived Performance as a Function of Actual Performance. Adapted from https://commons.wikimedia.org/wiki/File:Dunning-kruger_effect_-_percentile.svg.

11.5.1 Base Rate Fallacy

The *base rate fallacy* (aka base rate neglect) is the tendency to ignore information about the general probability of an event in favor of specific information about the event. The *base rate* is a marginal probability, which is the general probability of an event irrespective of other things. For example, the base rate of work-related injury is the general probability of experiencing a work-related injury, irrespective of other factors (e.g., the type of job, the person's age, the person's sex). Among the working population in the U.S., the lifetime prevalence of work-related injuries (i.e., the percent of people who will experience a work-related injury at some point in their lives), is ~35% (https://www.cdc.gov/mmwr/volumes/69/wr/mm6913a1.htm; archived at https://perma.cc/A2L6-WPEH). Thus, the base rate of work-related injuries in the U.S. is ~35%. The probability of work-related injuries is higher for some occupations (e.g., construction) and for some groups (e.g., men, 55–64-year-olds, Black or Multiracial, who are self-employed and have less than high school education) than others. Nevertheless, if we ignore all of the interacting factors, the general probability of work-related injuries is 35%. If we made a prediction that someone would be highly likely ($> 90\%$) to experience a work-related injury because they are male and self-employed, this would be ignoring the relatively lower base rate of work-related injury. Indeed, even men (36.7%) and self-employed individuals (41.2%) have less than a 50% chance of experiencing a work-related injury.

As applied to fantasy football, consider that you read about a potential sleeper Wide Receiver who had a stellar performance in a preseason game. If you select this player early on in the draft based on this information, this would be ignoring the general probability that most players who have a strong performance in the preseason do not perform as well in the regular season (i.e., base rate neglect). Performance in the preseason is not strongly predictive of performance in the regular season (https://fivethirtyeight.com/features/the-nfl-preseason-is-not-predictive-but-it-can-often-seem-that-way/; archived at https://perma.cc/FSG2-6AXE).

More information on base rates and how to counteract the base rate fallacy in described in Chapter 13.

11.5.2 Regression Fallacy

The regression fallacy is the failure to account for the fact that things tend to naturally fluctuate around their mean and that, after an extreme fluctuation, subsequent scores tend to regress (or reverse) to the mean. An example of the regression fallacy is the so-called Sports Illustrated or Madden cover jinx curse. The Sports Illustrated or Madden cover jinx curse is the urban legend that players who appear on the cover of

Sports Illustrated (the magazine) or Madden (the video game) will perform poorly. But, such a phenomenon can be more simply explained by regression to the mean (https://www.psychologytoday.com/us/blog/what-the-luck/201610/the-sports-illustrated-cover-jinx; archived at https://perma.cc/CZM9-TVFN). When a player has a superb season, they likely benefited from some degree to good luck, and it is unlikely that they will repeat such a stellar season the following year. Instead, they are likely—at least based on random fluctuation—to regress to their long-term mean.

Applied to fantasy football, consider that a Quarterback had a 5-touchdown game in Week 1. You are in need of a strong Quarterback, so you drop a solid player to pick him up. However, it is possible that the Quarterback benefited from playing against a week defense in the first game of the season. Future matchups may prove more difficult, and the player is unlikely to sustain such a solid performance consistently throughout the season (i.e., they are likely to regress toward their mean).

11.5.3 Hot Hand Fallacy

The "hot hand" is the idea that a player who experiences a successful outcome will have greater chance of success in subsequent attempts. For instance, in basketball, it is widely claimed by coaches, players, and commentators that players who have the hot hand (i.e., who are "on fire") are more likely to make shots because they made previous shots. Evidence on the hot hand is mixed. Considerable evidence historically has suggested that there is no such thing as a "hot hand" (Avugos et al., 2013; Bar-Eli et al., 2006; Gilovich et al., 1985). Some recent research, however, has suggested that there may be a small hot hand effect in some contexts (Bocskocsky et al., 2014; Miller & Sanjurjo, 2014). However, if any such effect exists, the hot hand may be limited to a small subset of players and the effect size of any hot hand effect appears to be small (Pelechrinis & Winston, 2022).

In football, when trying how to distribute the ball among multiple Running Backs, it is not uncommon to hear that a coach wants to give the ball to the Running Back with the "hot hand." In fantasy football, consider that a player just had a multiple touchdown game. Due to the hot hand fallacy, a manager might continue to start the player because they believe the player is "on fire" and is likely to continue to score at an unsustainable rate.

It is important consider whether such a string of strong performances are outliers and if the player may, in future games, regress to the mean. When considering whether strong performances are outliers and may regress to the mean, it is valuable to consider whether the player's health, skill, or situation has appreciably changed (compared to the player's earlier, weaker performances). Is the player finally fully healthy? Has the player appreciably improved in some skill that will benefit them in future games? Has the player's long-term

situation improved, such as moving up the depth chart, or receiving more carries/targets that is not tied to a specific opponent or game script? Or, alternatively, do the improvements appear to be driven by transient, game-specific factors, such as a the health of a teammate, the opponents they played, or the game script that ensued? If long-term outlook of the player has appreciably changed due to changes in the fundamentals of a player's value, such as their health, skill, or situation, it is less likely that such performance improvements will regress over the long run.

11.5.4 Sunk Cost Fallacy

A *sunk cost* is a cost (e.g., in money, time, or effort) that has already been incurred and cannot be recovered. For instance, if a person orders an expensive meal at a restaurant, the order is a sunk cost. The *sunk cost fallacy* is the tendency to continue an endeavor when there is a sunk cost. For instance, when ordering the expensive meal at the restaurant, a person may over-eat so that they feel that they eat their money's worth of food.

Applied to fantasy football, consider a situation in which you invest a lot of salary cap or a high draft pick to draft a promising player, but they repeatedly underperform. If you continue to start the player to justify your large investment, instead of benching him in favor of a higher-performing player, you are committing the sunk cost fallacy.

11.5.5 Gambler's Fallacy

The gambler's fallacy occurs due to an erroneous belief in the law of small numbers. The law of large numbers is a mathematical theorem that the average of a sufficiently large number of independent observations converges to the true value. For instance, if you flip a fair coin 1 million times, it is likely to land heads-up ~50% of the time. The law of *small* numbers (aka hasty generalization), by contrast, is an erroneous belief that small samples are representative of the populations from which they were drawn. For instance, if you flip a coin 10 times, belief in the law of small numbers would lead one to believe that the coin will flip heads-up exactly 5 times out of 10. However, in reality, the chance is less than 1 in 4 (24.6%) that exactly 5 of 10 coin flips turn up heads, as calculated below and as depicted in Figure 11.3:

```
dbinom(
  x = 5,      # number of coins that flip heads-up
  size = 10, # how many times you flip a coin
  prob = 0.5 # probability of a coin flipping heads-up (i.e., fair coin = 50%)
)
```

```
[1] 0.2460938
```

The `dbinom()` function in R provides the density of a binomial distribution. A binomial distribution is the probability of a particular number of successes (e.g., coins flipping heads-up) given a certain number of independent trials.

```
set.seed(52242)

numHeads <- rbinom(
  n = 100000,
  size = 10,
  prob = .5
)

simulationOfFlipping10Coins <- data.frame(
  numHeads = numHeads
)

simulationOfFlipping10Coins <- simulationOfFlipping10Coins %>%
  mutate(
    highlight = ifelse(numHeads == 5, "yes", "no")
  )

ggplot2::ggplot(
  data = simulationOfFlipping10Coins,
  mapping = aes(
    x = numHeads,
    fill = highlight)
) +
  geom_histogram(
    color = "#000000",
    bins = 11
  ) +
  scale_x_continuous(
    breaks = 0:10
  ) +
  scale_fill_manual(
    values = c(
      "yes" = "tomato",
      "no" = "gray")
  ) +
  labs(
    x = "Number of Coins Flipped Heads (out of 10 Coin Flips)",
    y = "Frequency",
    title = "Histogram of Number of Coins that Flip Heads-Up\nin a Simulation of 10 Coin Flips\n(w
```

```
  ) +
  theme_classic() +
  theme(legend.position = "none")
```

Figure 11.3 Histogram of Number of Coins that Flip Heads-Up in a Simulation of 10 Coin Flips (with 100,000 Replications).

Although 5 is the modal count of coins that flip heads-up out of 10 flips (i.e., it was more common than any other number), it is less common than the aggregate probability of flipping any number of heads besides 5. The probability of getting any other number of coin flips turning up heads (other than 5) is:

```
dbinom(
  x = c(0:4, 6:10),
  size = 10,
  prob = 0.5
) %>% sum()
```

```
[1] 0.7539063
```

The *gambler's fallacy* is the erroneous belief that future probabilities are influenced by past events, even when the events are independent. For example, a gambler may pay close attention to a particular slot machine. If the slot machine has not paid out in a while, the gambler may believe that the slot machine is about to pay out soon, and may start putting coins in the slots.

Applied to fantasy football, consider that a Quarterback has had several lousy games in a row. The gambler's fallacy might lead a manager to start the player under the belief that the player "is due" for a big game, expecting a strong performance from the player merely because they player has not had a good game in a while.

11.5.6 Conditional Probability Fallacy

We describe the conditional probability fallacy in Section 13.3.2 after introducing conditional probability in Section 13.3.1.3.

11.6 Conclusion

In conclusion, there are many heuristics, cognitive bias, and fallacies that people engage in when making judgments and predictions. It is prudent to be aware of these common biases and to work to counteract them. For instance, look for information that challenges or disconfirms your beliefs, and work to incorporate this information into your beliefs. Do your best to pursue observation, data analysis, and data interpretation with an open mind. You never know what important information you might discover if you go in with an open mind. Pay attention to fundamentals of a player's value, such as their health, skill, or situation when considering whether a player's performance may regress to the mean. If the strong performances appear to be driven by transient, game-specific factors, such as a the health of a teammate, the opponents they played, or the game script that ensued, future performances may be more likely to regress to the mean. In general, people tend to be overconfident in their predictions. There is considerable luck in fantasy football. Approach the task of prediction with humility; no one is consistently able to accurately predict how well players will perform.

12

Judgment Versus Actuarial Approaches to Prediction

12.1 Getting Started

12.1.1 Load Packages

12.2 Approaches to Prediction

There are two primary approaches to prediction: human judgment and the actuarial (i.e., statistical) method.

12.2.1 Human Judgment

Using the judgment method of prediction, all gathered information is collected and formulated into a prediction in the person's mind. The person selects, measures, and combines information and produces projections solely according to their experience and judgment. For instance, a proclaimed "fantasy expert" might use their experience, expertise, and judgment to make a prediction about how each player will perform by using whatever information and data they deem to be important, aggregating all of this information in their mind to make the prediction for each player.

12.2.2 Actuarial/Statistical Method

In the actuarial or statistical method of prediction, information is gathered and combined systematically in an evidence-based statistical prediction formula. The method is based on equations and data, so both are needed.

An example of a statistical method of prediction is the Violence Risk Appraisal Guide (Rice et al., 2013). The Violence Risk Appraisal Guide is used in an

attempt to predict violence and is used for parole decisions. For instance, the equation might be something like Equation 12.1:

$$\text{violence risk} = \beta \cdot \text{conduct disorder} + \beta \cdot \text{substance use} + \beta \cdot \text{suspended from school} + \beta \cdot \text{childhood aggression} + \ldots \tag{12.1}$$

Then, based on their score and the established cutoffs, a person is given a "low risk", "medium risk", or "high risk" designation.

An actuarial formula for projecting a Running Back's rushing yards might be something like Equation 12.2:

$$\text{rushing yards} = \beta \cdot \text{rushing yards last season} + \beta \cdot \text{age} + \beta \cdot \text{injury history} + \beta \cdot \text{strength of offensive line} + \ldots \tag{12.2}$$

The beta weights in the actuarial model reflect the relative weight to assign each predictor. For instance, in predicting rushing yards, a player's historical performance is likely the strongest predictor, whereas injury history might be a relatively weaker predictor. Thus, we might give historical performance a beta of 3 and injury history a beta of 1 to give a player's historical performance three times more weight than the player's injury history in predicting their rushing yards. For generating the actuarial model, you could obtain the beta weights for each predictor from multiple regression, from machine learning, or from prior research on the relative importance of each predictor.

12.2.3 Combining Human Judgment and Statistical Algorithms

There are numerous ways in which humans and statistical algorithms could be involved. On one extreme, humans make all judgments. On the other extreme, although humans may be involved in data collection, a statistical formula makes all decisions based on the input data, consistent with an actuarial approach. However, the human judgment and actuarial approaches can be combined in a hybrid way (Dana & Thomas, 2006). For example, to save time and money, a clinical psychologist might use an actuarial approach in all cases, but might only use a judgment approach when the actuarial approach gives a "positive" test. Or, the clinical psychologist might use both human judgment and an actuarial approach independently to see whether they agree. That is, the clinician may make a prediction based on their judgment and might also generate a prediction from an actuarial approach.

The challenge is what to do when the human and the algorithm disagree. Hypothetically, humans reviewing and adjusting the results from the statistical algorithm could lead to more accurate prediction. However, human input also could lead to the possibility or exacerbation of biased predictions. In general, with very few exceptions, actuarial approaches are as accurate or more

accurate than "expert" judgment (Ægisdóttir et al., 2006; Baird & Wagner, 2000; Dawes et al., 1989; Grove et al., 2000; Grove & Meehl, 1996). This is also likely true with respect to predicting player performance in sports (Den Hartigh et al., 2018). Moreover, the superiority of actuarial approaches to human judgment tends to hold even when the expert is given more information than the actuarial approach (Dawes et al., 1989). Allowing experts to override actuarial predictions consistently leads to lower predictive accuracy (Garb & Wood, 2019).

There is sometimes a misconception that formulas cannot account for qualitative information. However, that is not true. Qualitative information can be scored or coded to be quantified so that it can be included in statistical formulas. For instance, if an expert scout is able to meaningfully assess a player's cognitive and motivational factors (i.e., the "X factor" or "intangibles"), the scout can score this across multiple players and include these data in the actuarial prediction formula. For instance, the scout could use a rating scale (e.g., 1 = "poor"; 2 = "fair"; 3 = "good"; 4 = "very good"; 5 = "excellent") to code (i.e., translate) their qualitative judgment into a quantifiable rating that can be integrated with other information in the actuarial formula. That said, the quality of predictions rests on the quality and relevance of the assessment information for the particular prediction decision. If the assessment data are lousy, it is unlikely that a statistical algorithm (or a human for that matter) will make an accurate prediction: "Garbage in, garbage out". A statistical formula cannot rescue inaccurate assessment data.

12.3 Errors in Human Judgment

Human judgment is naturally subject to errors. Common heuristics, cognitive biases, and fallacies[1] are described in Chapter 11. Below, I describe a few errors to which human judgment seems particularly prone.

When operating freely, clinicians and medical experts (and humans more generally) tend to overestimate exceptions to the established rules (i.e., the broken leg syndrome). Meehl (1957) acknowledged that there may be some situations where it is glaringly obvious that the statistical formula would be incorrect because it fails to account for an important factor. He called these special cases "broken leg" cases, in which the human should deviate from the formula (i.e., broken leg countervailing). The example goes like this:

[1] sec-fallacies

> "If a sociologist were predicting whether Professor X would go to the movies on a certain night, he might have an equation involving age, academic specialty, and introversion score. The equation might yield a probability of .90 that Professor X goes to the movie tonight. But if the family doctor announced that Professor X had just broken his leg, no sensible sociologist would stick with the equation. Why didn't the factor of 'broken leg' appear in the formula? Because broken legs are very rare, and in the sociologist's entire sample of 500 criterion cases plus 250 cross-validating cases, he did not come upon a single instance of it. He uses the broken leg datum confidently, because 'broken leg' is a subclass of a larger class we may crudely denote as 'relatively immobilizing illness or injury,' and movie-attending is a subclass of a larger class of 'actions requiring moderate mobility.' " (Meehl, 1957, pp. 269–270)

However, people too often think that cases where they disagree with the statistical algorithm are broken leg cases. People too often think their case is an exception to the rule. As a result, they too often change the result of the statistical algorithm and are more likely to be wrong than right in doing so. Because actuarial methods are based on actual population levels (i.e., base rates), unique exceptions are not overestimated.

Actuarial predictions are perfectly reliable—they will always return the same conclusion given an identical set of data. The human judge is likely to both disagree with others and with themselves given the same set of symptoms.

The decision by an expert (all by all humans) is likely to be influenced by past experiences. Actuarial methods are based on objective algorithms, and past personal experience and personal biases do not factor into any decisions. Humans give weight to less relevant information, and often give too much weight to singular variables. Actuarial formulas do a better job of focusing on relevant variables. Computers are good at factoring in base rates. Humans ignore base rates (base rate neglect).

Computers are better at accurately weighing predictors and calculating unbiased risk estimates. In an actuarial formula, the relevant predictors are weighted according to their predictive power.

Humans are typically given no feedback on their judgments. To improve accuracy of judgments, it is important for feedback to be clear, consistent, and timely.

12.4 Humans Versus Computers

12.4.1 Advantages of Computers

Here are some advantages of computers over humans:

- Computers can process lots of information simultaneously. So can humans. But computers can to an even greater degree.
- Computers are faster at making calculations.
- Computations by computers are error-free (as long as the computations are programmed correctly).
- Computers' judgments will not be biased by fatigue or emotional responses.
- Computers' judgments will tend not to be biased in the way that humans' cognitive biases are. Computers are less likely to be overconfident in their judgments.
- Computers can more accurately weight the set of predictors based on large data sets. Humans tend to give too much weight to singular predictors.

12.4.2 Advantages of Humans

Computers are bad at some things too. Here are some advantages of humans over computers (as of now):

- Humans can be better at identifying patterns in data (but also can mistakenly identify patterns where there are none—i.e., illusory correlation).
- Humans can be flexible and take a different approach if a given approach is not working.
- Humans are better at tasks requiring creativity and imagination, such as developing theories that explain phenomena.
- Humans have the ability to reason, which is especially important when dealing with complex, abstract, or open-ended problems, or problems that have not been faced before (or for which we have insufficient data).
- Humans are better able to learn.
- Humans are better at holistic, gestalt processing, including facial and linguistic processing.

There *may* be situations in which a human judgment would do better than an actuarial judgment. One situation where human judgment would be important is when no actuarial method exists for the judgment or prediction. For instance, when no actuarial method exists for the diagnosis or disorder (e.g., suicide), it is up to the clinician. However, we could collect data on the

outcomes or on clinicians' judgments to develop an actuarial method that will be more reliable than the clinicians' judgments. That is, an actuarial method developed based on clinicians' judgments will be more accurate than clinicians' judgments. In other words, we do not necessarily need outcome data to develop an actuarial method. We could use the client's data as predictors of the clinicians' judgments to develop a structured approach to prediction that weighs factors similarly to clinicians, but with more reliable predictions.

Another situation in which human judgment could outperform a statistical algorithm is in true "broken leg" cases, e.g., important and rare events (edge cases) that are not yet accounted for by the algorithm.

Another situation in which human judgment could be preferable is if advanced, complex theories exist. Computers have a difficult time adhering to complex theories, so clinicians may be better suited. However, we do not have any of these complex theories in psychology that are accurate. We would need strong theory informed by data regarding causal influences, and accurate measures to assess them. However, no theories in psychology are that good. Nevertheless, predictive accuracy can be improved when considering theory (Garb & Wood, 2019; Silver, 2012).

If the prediction requires complex configural relations that a computer will have a difficult time replicating, a clinician's judgment may be preferred. Although the likelihood that a person can accurately work through these complex relations is theoretically possible, it is highly unlikely. Holistic pattern recognition (such as language and faces) tends to be better by humans than computers. But computers are getting better with holistic pattern recognition through machine learning.

In sum, the human seeks to integrate information to make a decision, but is biased.

12.4.3 Comparison of Evidence

Hundreds of studies have examined clinical versus actuarial prediction methods across many disciplines. Findings consistently show that actuarial methods are as accurate or more accurate than human judgment/prediction methods. "There is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly…as this one" (Meehl, 1986, pp. 373–374).

Actuarial methods are particularly valuable for criterion-referenced assessment tasks, in which the aim is to predict specific events or outcomes (Garb & Wood, 2019). For instance, actuarial methods have shown promise in predicting violence, criminal recidivism, psychosis onset, course of mental disorders, treatment selection, treatment failure, suicide attempts, and suicide (Garb & Wood, 2019).

Moreover, actuarial methods are explicit; they can be transparent and lead to informed scientific criticism to improve them. By contrast, human judgment methods are not typically transparent; human judgment relies on mental processes that are often difficult to specify.

12.5 Why Judgment is More Widely Used Than Statistical Formulas

Despite actuarial methods being generally more accurate than human judgment, judgment is much more widely used by clinicians. There are several reasons why actuarial methods have not caught on; one reason is professional traditions. Experts in any field do not like to think that a computer could outperform them. Some practitioners argue that judgment/prediction is an "art form" and that using a statistical formula is treating people like a number. However, using an approach (i.e., human judgment) that systematically leads to less accurate decisions and predictions is an ethical problem.

Some clinicians do not think that group averages (e.g., in terms of which treatment is most effective) apply to an individual client. This invokes the distinction between nomothetic (group-level) inferences and idiographic (individual-level) inferences. However, the scientific evidence and probability theory strongly indicate that it is better to generalize from group-level evidence than throwing out all the evidence and taking the approach of "anything goes." Clinicians frequently believe the broken leg fallacy, i.e., thinking that your client is an exception to the algorithmic prediction. In most cases, deviating from the statistical formula will result in less accurate predictions. People tend to overestimate the probability of low base rate conditions and events.

Another reason why actuarial methods have not caught on is the belief that receiving a treatment is the only thing that matters. But it is an empirical question which treatment is most effective for whom. What if we could do better? For example, we could potentially use a formula to identify the most effective treatment for a client. Some treatments are no better than placebo; other treatments are actually harmful (Lilienfeld, 2007; Williams et al., 2021).

Another reason why judgment methods are more widely used than actuarial methods is that so-called "experts" (and people in general) show overconfidence in their predictions—clinicians, experts, and humans in general think they are more accurate than they actually are. We see this when examining their calibration; their predictions tend to be miscalibrated. For example, things they report with 80% confidence occur less than 80% of the time, an example of overprecision in their predictions. Humans will sometimes be correct by chance, and they tend to mis-attribute that to their skill; humans tend

to remember the successes and forget the failures.

Another argument against using actuarial methods is that "no methods exist". In some cases, that is true—actuarial methods do not yet exist for some prediction problems. However, one can always create an algorithm of the experts' judgments, even if one does not have access to the outcome information. A model of clinicians' responses tends to be more accurate than clinicians' judgments themselves because the model gives the same outcome with the same input data—i.e., it is perfectly reliable.

Another argument from some clinicians is that, "My job is to understand, not to predict". But what kind of understanding does not involve predictions? Accurate predictions help in understanding. Knowing how people would perform in different conditions is the same thing as good understanding.

12.6 Best Actuarial Approaches to Prediction

The best actuarial models tend to be relatively simple (parsimonious), that can account for one or several of the most important predictors and their optimal weightings, and that account for the base rate of the phenomenon. Even unit-weighted formulas (formulas whose predictor variables are equally weighted with a weight of one) can sometimes generalize better to other samples than complex weightings (Garb & Wood, 2019). Differential weightings sometimes capture random variance and over-fit the model, thus leading to predictive accuracy shrinkage in cross-validation samples (Garb & Wood, 2019), as described below. The choice of predictor variables often matters more than their weighting.

In general, there is often shrinkage of estimates from training data set to a test data set. *Shrinkage* is when variables with stronger predictive power in the original data set tend to show somewhat smaller predictive power (smaller regression coefficients) when applied to new groups. Shrinkage reflects a model overfitting (i.e., fitting to error by capitalizing on chance). Shrinkage is especially likely when the original sample is small and/or unrepresentative and the number of variables considered for inclusion is large. Cross-validation with large, representative samples can help evaluate the amount of shrinkage of estimates, particularly for more complex models such as machine learning models (Ursenbach et al., 2019). Ideally, cross-validation would be conducted with a separate sample (external cross-validation) to see the generalizability of estimates. However, you can also do internal cross-validation. For example, you can perform k-fold cross-validation, where you:

- split the data set into k groups

- for each unique group:
  - take the group as a hold-out data set (also called a test data set)
  - take the remaining groups as a training data set
  - fit a model on the training data set and evaluate it on the test data set
  - after all k-folds have been used as the test data set, and all models have been fit, you average the estimates across the models, which presumably yields more robust, generalizable estimates

An emerging technique that holds promise for increasing predictive accuracy of actuarial methods is machine learning (Garb & Wood, 2019). However, one challenge of some machine learning techniques is that they are like a "black box" and are not transparent, which raises ethical concerns (Garb & Wood, 2019). machine learning may be most valuable when the data available are complex and there are many predictor variables (Garb & Wood, 2019).

12.7 Conclusion

In general, it is better to develop and use structured, actuarial approaches than informal approaches that rely on human judgment or judgment by "so-called" experts. Actuarial approaches to prediction tend to be as accurate or more accurate than expert judgment. Nevertheless, in many domains, human judgment tends to be much more widely used than actuarial approaches.

13

Base Rates

13.1 Getting Started

13.1.1 Load Packages

```
library("petersenlab")
```

13.2 Overview

Predicting player performance is a complex prediction task. Performance is probabilistically influenced by many processes, including processes internal to the player in addition to external processes. Moreover, people's performance occurs in the context of a dynamic system with nonlinear, probabilistic, and cascading influences that change across time. The ever-changing system makes behavior challenging to predict. And, similar to chaos theory, one small change in the system can lead to large differences later on. Moreover, there are important factors to keep in mind when making predictions.

Let's consider a prediction example, assuming the following probabilities:

- The probability of contracting HIV is .3%
- The probability of a positive test for HIV is 1%
- The probability of a positive test if you have HIV is 95%

What is the probability of HIV if you have a positive test?

As we will see, the probability is: $\frac{95\% \times .3\%}{1\%} = 28.5\%$. So based on the above probabilities, if you have a positive test, the probability that you have HIV is 28.5%. Most people tend to vastly overestimate the likelihood that the person has HIV in this example. Why? Because they do not pay enough attention to the base rate (in this example, the base rate of HIV is .3%).

13.3 Issues Around Probability

13.3.1 Types of Probabilities

It is important to distinguish between different types of probabilities: marginal probabilities, joint probabilities, and conditional probabilities.

13.3.1.1 Base Rate (Marginal Probability)

The *base rate* is a marginal probability, which is the general probability of an event irrespective of other things. For instance, the base rate of HIV is the probability of developing HIV. In the U.S., the prevalence rate of HIV is ~0.4% of the adult population[1] (archived at https://perma.cc/8GE6-GAPC).

For instance, we can consider the following marginal probabilities:

$P(C_i)$ is the probability (i.e., base rate) of a classification, C, independent of other things. A base rate is often used as the "*prior probability*" in a Bayesian model. In our example above, $P(C_i)$ is the base rate (i.e., prevalence) of HIV in the population: $P(\text{HIV}) = .3\%$. $P(R_i)$ is the probability (base rate) of a response, R, independent of other things. In the example above, $P(R_i)$ is the base rate of a positive test for HIV: $P(\text{positive test}) = 1\%$. The base rate of a positive test is known as the *positivity rate* or *selection ratio*.

13.3.1.2 Joint Probability

A *joint probability* is the probability of two (or more) events occurring simultaneously. For instance, the probability of events A and B both occurring together is $P(A, B)$. A joint probability can be calculated using the marginal probability of each event, as in Equation 13.1:

$$P(A, B) = P(A) \cdot P(B) \tag{13.1}$$

Conversely (and rearranging the terms for the calculation of conditional probability), a joint probability can also be calculated using the conditional probability and marginal probability, as in Equation 13.2:

$$P(A, B) = P(A|B) \cdot P(B) \tag{13.2}$$

[1] https://map.aidsvu.org/profiles/nation/usa/overview

13.3.1.3 Conditional Probability

A *conditional probability* is the probability of one event occurring given the occurrence of another event. Conditional probabilities are written as: $P(A|B)$. This is read as the probability that event A occurs given that event B occurred. For instance, we can consider the following conditional probabilities:

$P(C|R)$ is the probability of a classification, C, given a response, R. In other words, $P(C|R)$ is the probability of having HIV given a positive test: $P(\text{HIV}|\text{positive test})$. $P(R|C)$ is the probability of a response, R, given a classification, C. In the example above, $P(R|C)$ is the probability of having a positive test given that a person has HIV: $P(\text{positive test}|\text{HIV}) = 95\%$.

A conditional probability can be calculated using the joint probability and marginal probability (base rate), as in Equation 13.3:

$$P(A, B) = P(A|B) \cdot P(B) \tag{13.3}$$

13.3.2 Confusion of the Inverse

A conditional probability is not the same thing as its reverse (or inverse) conditional probability. Unless the base rate of the two events (C and R) are the same, $P(C|R) \neq P(R|C)$. However, people frequently make the mistake of thinking that two inverse conditional probabilities are the same. This mistake is known as the "confusion of the inverse", or the "inverse fallacy", or the "conditional probability fallacy". The confusion of inverse probabilities is the logical error of representative thinking that leads people to assume that the probability of C given R is the same as the probability of R given C, even though this is not true. As a few examples to demonstrate the logical fallacy, if 93% of breast cancers occur in high-risk women, this does not mean that 93% of high-risk women will eventually get breast cancer. As another example, if 77% of car accidents take place within 15 miles of a driver's home, this does not mean that you will get in an accident 77% of times you drive within 15 miles of your home.

Which car is the most frequently stolen? It is often the Honda Accord or Honda Civic—probably because they are among the most popular/commonly available cars. The probability that the car is a Honda Accord given that a car was stolen (p(Honda Accord | Stolen)) is what the media reports and what the police care about. However, that is not what buyers and car insurance companies should care about. Instead, they care about the probability that the car will be stolen given that it is a Honda Accord (p(Stolen | Honda Accord)).

Applied to fantasy football, the probability that a given player will be injured given that he is a Running Back (p(Injured | RB)) is not the same as the

probability that a given player is a Running Back given that he is injured (p(RB | Injured)).

13.3.3 Bayes' Theorem

An alternative way of calculating a conditional probability is using the inverse conditional probability (instead of the joint probability). This is known as Bayes' theorem. Bayes' theorem can help us calculate a conditional probability of some classification, C, given some response, R, if we know the inverse conditional probability and the base rate (marginal probability) of each. Bayes' theorem is in Equation 13.4:

$$P(C|R) = \frac{P(R|C) \cdot P(C_i)}{P(R_i)} \tag{13.4}$$

Or, equivalently (rearranging the terms):

$$\frac{P(C|R)}{P(R|C)} = \frac{P(C_i)}{P(R_i)} \tag{13.5}$$

Or, equivalently (rearranging the terms):

$$\frac{P(C|R)}{P(C_i)} = \frac{P(R|C)}{P(R_i)} \tag{13.6}$$

More generally, Bayes' theorem has been described as:

$$\begin{aligned} P(H|E) &= \frac{P(E|H) \cdot P(H)}{P(E)} \\ \text{posterior probability} &= \frac{\text{likelihood} \times \text{prior probability}}{\text{model evidence}} \end{aligned} \tag{13.7}$$

where H is the hypothesis, and E is the evidence—the new information that was not used in computing the prior probability.

In Bayesian terms, the *posterior probability* is the conditional probability of one event occurring given another event—it is the updated probability after the evidence is considered. In this case, the posterior probability is the probability of the classification occurring (C) given the response (R). The *likelihood* is the inverse conditional probability—the probability of the response (R) occurring given the classification (C). The *prior probability* is the marginal probability of the event (i.e., the classification) occurring, before we take into account any new information. The *model evidence* is the marginal probability of the other event occurring—i.e., the marginal probability of seeing the evidence.

Bayes' theorem provides the foundation for a paradigm of statistics called Bayesian statistics, which (unlike frequentist statistics) does not use p-values.

In the HIV example above, we can calculate the conditional probability of HIV given a positive test using three terms: the conditional probability of a positive test given HIV (i.e., the sensitivity of the test), the base rate of HIV, and the base rate of a positive test for HIV. The conditional probability of HIV given a positive test is in Equation 13.8:

$$
\begin{aligned}
P(C|R) &= \frac{P(R|C) \cdot P(C_i)}{P(R_i)} \\
P(\text{HIV}|\text{positive test}) &= \frac{P(\text{positive test}|\text{HIV}) \cdot P(\text{HIV})}{P(\text{positive test})} \\
&= \frac{\text{sensitivity of test} \times \text{base rate of HIV}}{\text{base rate of positive test}} \\
&= \frac{95\% \times .3\%}{1\%} = \frac{.95 \times .003}{.01} \\
&= 28.5\%
\end{aligned}
\tag{13.8}
$$

The `petersenlab`[2] package (Petersen, 2024a) contains the `pAgivenB()` function that estimates the probability of one event, A, given another event, B.

```
petersenlab::pAgivenB(
  pBgivenA = .95,
  pA = .003,
  pB = .01)
```

```
[1] 0.285
```

Thus, assuming the probabilities in the example above, the conditional probability of having HIV if a person has a positive test is 28.5%. Given a positive test, chances are higher than not that the person does not have HIV.

Now let's see what happens if the person tests positive a second time. We would revise our "prior probability" for HIV from the general prevalence in the population (0.3%) to be the "posterior probability" of HIV given a first positive test (28.5%). This is known as *Bayesian updating*. We would also update the "evidence" to be the marginal probability of getting a second positive test.

If we do not know a marginal probability (i.e., base rate) of an event (e.g., getting a second positive test), we can calculate a marginal probability with the *law of total probability* using conditional probabilities and the marginal

[2] https://cran.r-project.org/web/packages/petersenlab/index.html

probability of another event (e.g., having HIV). According to the law of total probability, the probability of getting a positive test is the probability that a person with HIV gets a positive test (i.e., sensitivity) times the base rate of HIV plus the probability that a person without HIV gets a positive test (i.e., false positive rate) times the base rate of not having HIV, as in Equation 13.9:

$$\begin{aligned} P(\text{not } C_i) &= 1 - P(C_i) \\ P(R_i) &= P(R|C) \cdot P(C_i) + P(R|\text{not } C) \cdot P(\text{not } C_i) \\ 1\% &= 95\% \times .3\% + P(R|\text{not } C) \times 99.7\% \end{aligned} \tag{13.9}$$

In this case, we know the marginal probability ($P(R_i)$), and we can use that to solve for the unknown conditional probability that reflects the false positive rate ($P(R|\text{not } C)$), as in Equation 13.10:

$$\begin{aligned} P(R_i) &= P(R|C) \cdot P(C_i) + P(R|\text{not } C) \cdot P(\text{not } C_i) \\ P(R_i) - [P(R|\text{not } C) \cdot P(\text{not } C_i)] &= P(R|C) \cdot P(C_i) && \text{Move } P(R|\text{not } C) \text{ to the left side} \\ -[P(R|\text{not } C) \cdot P(\text{not } C_i)] &= P(R|C) \cdot P(C_i) - P(R_i) && \text{Move } P(R_i) \text{ to the right side} \\ P(R|\text{not } C) \cdot P(\text{not } C_i) &= P(R_i) - [P(R|C) \cdot P(C_i)] && \text{Multiply by } -1 \\ P(R|\text{not } C) &= \frac{P(R_i) - [P(R|C) \cdot P(C_i)]}{P(\text{not } C_i)} && \text{Divide by } P(R|\text{not } C) \\ &= \frac{1\% - [95\% \times .3\%]}{99.7\%} = \frac{.01 - [.95 \times .003]}{.997} \\ &= .7171515\% \end{aligned} \tag{13.10}$$

The `petersenlab`[3] package (Petersen, 2024a) contains the `pBgivenNotA()` function that estimates the probability of one event, B, given that another event, A, did not occur.

```
petersenlab::pBgivenNotA(
  pBgivenA = .95,
  pA = .003,
  pB = .01)
```

```
[1] 0.007171515
```

With this conditional probability ($P(R|\text{not } C)$), the updated marginal probability of having HIV ($P(C_i)$), and the updated marginal probability of not having HIV ($P(\text{not } C_i)$), we can now calculate an updated estimate of the marginal probability of getting a second positive test. The probability of getting a second positive test is the probability that a person with HIV gets a second positive test (i.e., sensitivity) times the updated probability of HIV plus the probability that a person without HIV gets a second positive test (i.e., false positive rate) times the updated probability of not having HIV, as in Equation 13.11:

[3] https://cran.r-project.org/web/packages/petersenlab/index.html

$$
\begin{aligned}
P(R_i) &= P(R|C) \cdot P(C_i) + P(R|\text{not } C) \cdot P(\text{not } C_i) \\
&= 95\% \times 28.5\% + .7171515\% \times 71.5\% = .95 \times .285 + .007171515 \times .715 \\
&= 27.58776\%
\end{aligned}
\tag{13.11}
$$

The `petersenlab`[4] package (Petersen, 2024a) contains the `pB()` function that estimates the marginal probability of one event, B.

```
petersenlab::pB(
  pBgivenA = .95,
  pA = .285,
  pBgivenNotA = .007171515)
```

```
[1] 0.2758776
```

We then substitute the updated marginal probability of HIV ($P(C_i)$) and the updated marginal probability of getting a second positive test ($P(R_i)$) into Bayes' theorem to get the probability that the person has HIV if they have a second positive test (assuming the errors of each test are independent, i.e., uncorrelated), as in Equation 13.12:

$$
\begin{aligned}
P(C|R) &= \frac{P(R|C) \cdot P(C_i)}{P(R_i)} \\
P(\text{HIV}|\text{a second positive test}) &= \frac{P(\text{a second positive test}|\text{HIV}) \cdot P(\text{HIV})}{P(\text{a second positive test})} \\
&= \frac{\text{sensitivity of test} \times \text{updated base rate of HIV}}{\text{updated base rate of positive test}} \\
&= \frac{95\% \times 28.5\%}{27.58776\%} \\
&= 98.14\%
\end{aligned}
\tag{13.12}
$$

The `petersenlab`[5] package (Petersen, 2024a) contains the `pAgivenB()` function that estimates the probability of one event, A, given another event, B.

```
petersenlab::pAgivenB(
  pBgivenA = .95,
  pA = .285,
  pB = .2758776)
```

```
[1] 0.9814135
```

[4] https://cran.r-project.org/web/packages/petersenlab/index.html
[5] https://github.com/DevPsyLab/petersenlab

Thus, a second positive test greatly increases the posterior probability that the person has HIV from 28.5% to over 98%.

As seen in the rearranged formula in Equation 13.5, the ratio of the conditional probabilities is equal to the ratio of the base rates. Thus, it is important to consider base rates. People have a strong tendency to ignore (or give insufficient weight to) base rates when making predictions. The failure to consider the base rate when making predictions when given specific information about a case is known as the base rate fallacy or as base rate neglect. For example, people tend to say that the probability of a rare event is more likely than it actually is given specific information.

As seen in the rearranged formula in Equation 13.6, the inverse conditional probabilities ($P(C|R)$ and $P(R|C)$) are not equal unless the base rates of C and R are the same. If the base rates are not equal, we are making at least some prediction errors. If $P(C_i) > P(R_i)$, our predictions must include some false negatives. If $P(R_i) > P(C_i)$, our predictions must include some false positives.

In sum, the marginal probability, including the prior probability or base rate, should be weighed heavily in predictions unless there are sufficient data to indicate otherwise, i.e., to update the posterior probability based on new evidence. Bayes' theorem provides a powerful tool to anchor predictions to the base rate unless sufficient evidence changes the posterior probability (by updating the evidence and prior probability).

13.4 Base Rate of Rookie Performance

13.4.1 Quarterbacks

13.4.2 Running Backs

13.5 How to Account for Base Rates

There are various ways to account for base rates, including the use of actuarial formulas and the use of Bayesian updating.

13.5.1 Actuarial Formula

One approach to account for base rates is to use actuarial formulas (rather than human judgment) to make the predictions. Actuarial formulas based on multiple regression or machine learning can account for the base rate of the event.

13.5.2 Bayesian Updating

Another approach to account for base rates is to leverage Bayes' theorem, using Bayesian updating and the probability nomogram. Bayesian updating is a form of anchoring and adjustment; however, unlike the anchoring and adjustment heuristic, it is a systematic approach to anchoring and adjustment that anchors one's predictions to the base rate, and then adjusts according to new information. That is, we start with a pretest probability (i.e., base rate) and update our predictions based on the extent of new information (i.e., the likelihood ratio).

To perform Bayesian updating involves comparing the relative probability of two outcomes, $P(C|R)$ versus $P(\text{not } C|R)$. If we want to compare the relative probability of two outcomes, we can use the odds form of Bayes' theorem, as in Equation 13.13:

$$\begin{aligned}
P(C|R) &= \frac{P(R|C) \cdot P(C_i)}{P(R_i)} \\
P(\text{not } C|R) &= \frac{P(R|\text{not } C) \cdot P(\text{not } C_i)}{P(R_i)} \\
\frac{P(C|R)}{P(\text{not } C|R)} &= \frac{\frac{P(R|C) \cdot P(C_i)}{P(R_i)}}{\frac{P(R|\text{not } C) \cdot P(\text{not } C_i)}{P(R_i)}} \\
&= \frac{P(R|C) \cdot P(C_i)}{P(R|\text{not } C) \cdot P(\text{not } C_i)} \\
&= \frac{P(C_i)}{P(\text{not } C_i)} \times \frac{P(R|C)}{P(R|\text{not } C)} \\
\text{posterior odds} &= \text{prior odds} \times \text{likelihood ratio}
\end{aligned} \tag{13.13}$$

As presented in Equation 13.13, the posttest (or posterior) odds are equal to the pretest odds multiplied by the likelihood ratio. Below, we describe the likelihood ratio.

13.5.2.1 Diagnostic Likelihood Ratio

A likelihood ratio is the ratio of two probabilities. It can be used to compare the likelihood of two possibilities. The diagnostic likelihood ratio is an index

of the predictive validity of an instrument: it is the ratio of the probability that a test result is correct to the probability that the test result is incorrect. The diagnostic likelihood ratio is also called the risk ratio. There are two types of diagnostic likelihood ratios: the positive likelihood ratio and the negative likelihood ratio.

13.5.2.1.1 Positive Likelihood Ratio (LR+)

The positive likelihood ratio (LR+) compares the true positive rate to the false positive rate. Positive likelihood ratio values range from 1 to infinity. Higher values reflect greater accuracy, because it indicates the degree to which a true positive is more likely than a false positive. The formula for calculating the positive likelihood ratio is in Equation 13.14.

$$\begin{aligned}\text{positive likelihood ratio (LR+)} &= \frac{\text{TPR}}{\text{FPR}} \\ &= \frac{P(R|C)}{P(R|\text{not } C)} \\ &= \frac{P(R|C)}{1 - P(\text{not } R|\text{not } C)} \\ &= \frac{\text{sensitivity}}{1 - \text{specificity}}\end{aligned} \tag{13.14}$$

13.5.2.1.2 Negative Likelihood Ratio (LR−)

The negative likelihood ratio (LR−) compares the false negative rate to the true negative rate. Negative likelihood ratio values range from 0 to 1. Smaller values reflect greater accuracy, because it indicates that a false negative is less likely than a true negative. The formula for calculating the negative likelihood ratio is in Equation 13.15.

$$\begin{aligned}\text{negative likelihood ratio (LR−)} &= \frac{\text{FNR}}{\text{TNR}} \\ &= \frac{P(\text{not } R|C)}{P(\text{not } R|\text{not } C)} \\ &= \frac{1 - P(R|C)}{P(\text{not } R|\text{not } C)} \\ &= \frac{1 - \text{sensitivity}}{\text{specificity}}\end{aligned} \tag{13.15}$$

13.5.2.2 Probability Nomogram

Using Bayes' theorem (described in Section 13.3.3), solving for posttest odds (based on pretest odds and the likelihood ratio, as in Equation 13.13), and converting odds to probabilities, we can use a Fagan probability nomogram to determine the posttest probability following a test result. The calculation of posttest probability is described in INSERT. A *probability nomogram* is a way of visually applying Bayes' theorem to determine the posttest probability of having a condition based on the pretest (or prior) probability and likelihood ratio, as depicted in Figure 13.1. To use a probability nomogram, connect the dots from the starting probability (left line) with the likelihood ratio (middle line) to see the updated probability. The updated (posttest) probability is where the connecting line crosses the third, right line.

Figure 13.1 Probability Nomogram. (Figure retrieved from https://upload.wikimedia.org/wikipedia/commons/thumb/6/66/Fagan_nomogram.svg/945px-Fagan_nomogram.svg.png).

For instance, if the starting probability is 0.5% and the likelihood ratio is 10 (e.g., sensitivity = .90, specificity = .91: likelihood ratio = $\frac{\text{sensitivity}}{1-\text{specificity}} = \frac{.9}{1-.91} = 10$) from a positive test (i.e., positive likelihood ratio), the updated probability is less than 5%, as depicted in Figure 13.2. The `petersenlab`[6] package (Petersen, 2024a) contains the `posttestProbability()` function that

[6] https://github.com/DevPsyLab/petersenlab

estimates the posttest probability of an event, given the pretest probability and the likelihood ratio, or given the pretest probability and the sensitivity (SN) and specificity (SP) of the test.

```
petersenlab::posttestProbability(
  pretestProb = .005,
  likelihoodRatio = 10)
```

```
[1] 0.04784689
```

```
petersenlab::posttestProbability(
  pretestProb = .005,
  SN = .90,
  SP = .91)
```

```
[1] 0.04784689
```

The function can also estimate the posttest probability of an event given the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

```
petersenlab::posttestProbability(
  TP = 450,
  TN = 90545,
  FP = 8955,
  FN = 50)
```

```
[1] 0.04784689
```

We discuss true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity (SN), and specificity (SP) in INSERT.

If the starting probability is 0.5% and the likelihood ratio is 0.11 from a negative test (i.e., negative likelihood ratio), the updated probability is nearly indistinguishable from zero (0.05%).

```
petersenlab::posttestProbability(
  pretestProb = .005,
  likelihoodRatio = 0.11)
```

```
[1] 0.0005524584
```

Figure 13.2 Probability Nomogram Example. (Figure adapted from https://upload.wikimedia.org/wikipedia/commons/thumb/6/66/Fagan_nomogram.svg/945px-Fagan_nomogram.svg.png. Also provided in: Petersen (2024b) and Petersen (2024c).)

A probability nomogram calculator can be found at the following link: http://araw.mede.uic.edu/cgi-bin/testcalc.pl (archived at https://perma.cc/X8TF-7YBX). The `petersenlab`[7] package (Petersen, 2024a) contains the `nomogrammer()` function that creates a nomogram plot using the positive and negative likelihood ratio or using the sensitivity (SN) and specificity (SP) of the test, as adapted from Adam Chekroud (https://github.com/achekroud/nomogrammer):

```
petersenlab::nomogrammer(
  pretestProb = .005,
  SN = 0.90,
  SP = 0.91)
```

[7] https://github.com/DevPsyLab/petersenlab

```
petersenlab::nomogrammer(
  pretestProb = .005,
  PLR = 10,
  NLR = 0.11)
```

The blue line indicates the posterior probability of the condition given a positive test. The pink line indicates the posterior probability of the condition given a negative test.

The function can also create a nomogram plot using the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

```
petersenlab::nomogrammer(
  TP = 450,
  TN = 90545,
  FP = 8955,
  FN = 50)
```

13.6 Conclusion

Fantasy performance—and behavior more generally—is challenging to predict. People commonly demonstrate biases and fallacies when making predictions. People tend to ignore base rates (base rate fallacy) when making predictions. They also tend to confuse inverse conditional probabilities (conditional probability fallacy). Bayes' theorem provides a way to convert from one conditional probability to its inverse conditional probability using the base rate of each event. There are various ways to account for base rates for more accurate predictions, including through the use of actuarial formulas and Bayesian updating. Bayesian updating uses Bayes' theorem to calculate a posttest probability from a pretest probability and a test result (likelihood ratio). The probability nomogram is a visual approach to Bayesian updating.

14

Evaluation of Prediction/Forecasting Accuracy

14.1 Getting Started

14.1.1 Load Packages

```
library("petersenlab")
library("tidyverse")
library("pROC")
library("magrittr")
library("viridis")
```

14.2 Overview

Predictions can come in different types. Some predictions involve categorical data, whereas other predictions involve continuous data. When dealing with a dichotomous (nominal data that are binary) predictor and outcome variable (or continuous data that have been dichotomized using a cutoff), we can evaluate predictions using a 2x2 table known as a confusion matrix (see INSERT), or with logistic regression models. When dealing with a continuous outcome variable (e.g., ordinal, interval, or ratio data), we can evaluate predictions using multiple regression or similar variants such as structural equation modeling and mixed models.

In fantasy football, we most commonly predict continuous outcome variables (e.g., fantasy points, rushing yards). Nevertheless, it is also important to understand principles in the prediction of categorical outcomes variables.

In any domain, it is important to evaluate the accuracy of predictions, so we

can know how (in)accurate we are, and we can strive to continually improve our predictions. Fantasy performance—and human behavior more general—is incredibly challenging to predict. In fantasy football, there is considerable luck/chance/randomness. There are relatively few (i.e. 17) games, and there is a sizeable injury risk for each player in a given game. These and other factors combine to render fantasy football predictions not highly accurate. But, first, let's learn about the various ways we can evaluate the accuracy of predictions.

14.3 Types of Accuracy

There are two primary dimensions of accuracy: (1) discrimination and (2) calibration. Discrimination and calibration are distinct forms of accuracy. Just because predictions are high in one form of accuracy does not mean that they will be high in the other form of accuracy. As described by Lindhiem et al. (2020), predictions can follow any of the following configurations (and anywhere in between):

- high discrimination, high calibration
- high discrimination, low calibration
- low discrimination, high calibration
- low discrimination, low calibration

Some general indexes of accuracy combine discrimination and calibration, as described in Section 14.3.3.

In addition, accuracy indices can be threshold-dependent or -independent and can be scale-dependent or -independent. Threshold-dependent accuracy indices differ based on the cutoff (i.e., threshold), whereas threshold-independent accuracy indices do not. Thus, raising or lowering the cutoff will change threshold-dependent accuracy indices. Scale-dependent accuracy indices depend on the metric/scale of the data, whereas scale-independent accuracy indices do not. Thus, scale-dependent accuracy indices cannot be directly compared when using measures of differing scales, whereas scale-independent accuracy indices can be compared across data of differing scales.

14.3.1 Discrimination

When dealing with a categorical outcome, discrimination is the ability to separate events from non-events. When dealing with a continuous outcome, discrimination is the strength of the association between the predictor and

the outcome. Aspects of discrimination at a particular cutoff (e.g., sensitivity, specificity, area under the ROC curve) are described in INSERT.

14.3.2 Calibration

When dealing with a categorical outcome, calibration is the degree to which a probabilistic estimate of an event reflects the true underlying probability of the event. When dealing with a continuous outcome, calibration is the degree to which the predicted values are close in value to the outcome values. The importance of examining calibration (in addition to discrimination) is described by Lindhiem et al. (2020).

Calibration is relevant to all kinds of predictions, including weather forecasts. For instance, on the days that the meteorologist says there is a 60% chance of rain, it should rain about 60% of the time. Calibration is also important for fantasy football predictions. When projections state that a group of players is each expected to score 200 points, their projections would be miscalibrated if those players scored only 150 points on average.

There are four general patterns of miscalibration: overextremity, underextremity, overprediction, and underprediction (see Figure 14.7). *Overextremity* exists when the predicted probabilites are too close to the extremes (zero or one). *Underextremity* exists when the predicted probabilities are too far away from the extremes. *Overprediction* exists when the predicted probabilities are consistently greater than the observed probabilities. *Underprediction* exists when the predicted probabilities are consistently less than the observed probabilities. For a more thorough description of these types of miscalibration, see Lindhiem et al. (2020).

Indices for evaluating calibration are described in Section 14.7.3.

14.3.3 General Accuracy

General accuracy indices combine estimates of discrimination and calibration.

14.4 Prediction of Categorical Outcomes

To evaluate the accuracy of our predictions for categorical outcome variables (e.g., binary, dichotomous, or nominal data), we can use either threshold-dependent or threshold-independent accuracy indices.

14.5 Prediction of Continuous Outcomes

To evaluate the accuracy of our predictions for continuous outcome variables (e.g., ordinal, interval, or ratio data), the outcome variable does not have cutoffs, so we would use threshold-independent accuracy indices.

14.6 Threshold-Dependent Accuracy Indices

14.6.1 Decision Outcomes

To consider how we can evaluate the accuracy of predictions for a categorical outcome, consider an example adapted from Meehl & Rosen (1955). The military conducts a test of its prospective members to screen out applicants who would likely fail basic training. To evaluate the accuracy of our predictions using the test, we can examine a confusion matrix. A confusion matrix is a matrix that presents the predicted outcome on one dimension and the actual outcome (truth) on the other dimension. If the predictions and outcomes are dichotomous, the confusion matrix is a 2x2 matrix with two rows and two columns that represent four possible predicted-actual combinations (decision outcomes): true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

When discussing the four decision outcomes, "true" means an accurate judgment, whereas "false" means an inaccurate judgment. "Positive" means that the judgment was that the person has the characteristic of interest, whereas "negative" means that the judgment was that the person does not have the characteristic of interest. A *true positive* is a correct judgment (or prediction) where the judgment was that the person has (or will have) the characteristic of interest, and, in truth, they actually have (or will have) the characteristic. A *true negative* is a correct judgment (or prediction) where the judgment was that the person does not have (or will not have) the characteristic of interest, and, in truth, they actually do not have (or will not have) the characteristic. A *false positive* is an incorrect judgment (or prediction) where the judgment was that the person has (or will have) the characteristic of interest, and, in truth, they actually do not have (or will not have) the characteristic. A *false negative* is an incorrect judgment (or prediction) where the judgment was that the person does not have (or will not have) the characteristic of interest, and, in truth, they actually do have (or will have) the characteristic.

An example of a confusion matrix is in INSERT.

With the information in the confusion matrix, we can calculate the marginal sums and the proportion of people in each cell (in parentheses), as depicted in INSERT.

That is, we can sum across the rows and columns to identify how many people actually showed poor adjustment ($n = 100$) versus good adjustment ($n = 1,900$), and how many people were selected to reject ($n = 508$) versus retain ($n = 1,492$). If we sum the column of predicted marginal sums ($508 + 1,492$) or the row of actual marginal sums ($100 + 1,900$), we get the total number of people ($N = 2,000$).

Based on the marginal sums, we can compute the marginal probabilities, as depicted in INSERT.

The marginal probability of the person having the characteristic of interest (i.e., showing poor adjustment) is called the *base rate* (BR). That is, the base rate is the proportion of people who have the characteristic. It is calculated by dividing the number of people with poor adjustment ($n = 100$) by the total number of people ($N = 2,000$): $BR = \frac{FN+TP}{N}$. Here, the base rate reflects the prevalence of poor adjustment. In this case, the base rate is .05, so there is a 5% chance that an applicant will be poorly adjusted. The marginal probability of good adjustment is equal to 1 minus the base rate of poor adjustment.

The marginal probability of predicting that a person has the characteristic (i.e., rejecting a person) is called the *selection ratio* (SR). The selection ratio is the proportion of people who will be selected (in this case, rejected rather than retained); i.e., the proportion of people who are identified as having the characteristic. The selection ratio is calculated by dividing the number of people selected to reject ($n = 508$) by the total number of people ($N = 2,000$): $SR = \frac{TP+FP}{N}$. In this case, the selection ratio is .25, so 25% of people are rejected. The marginal probability of not selecting someone to reject (i.e., the marginal probability of retaining) is equal to 1 minus the selection ratio.

The selection ratio might be something that the test dictates according to its cutoff score. Or, the selection ratio might be imposed by external factors that place limits on how many people you can assign a positive test value. For instance, when deciding whether to treat a client, the selection ratio may depend on how many therapists are available and how many cases can be treated.

14.6.2 Percent Accuracy

Based on the confusion matrix, we can calculate the prediction accuracy based on the percent accuracy of the predictions. The percent accuracy is the number of correct predictions divided by the total number of predictions, and multiplied by 100. In the context of a confusion matrix, this is calculated as:

$100\% \times \frac{\text{TP+TN}}{N}$. In this case, our percent accuracy was 78%—that is, 78% of our predictions were accurate, and 22% of our predictions were inaccurate.

14.6.3 Percent Accuracy by Chance

78% sounds pretty accurate. And it is much higher than 50%, so we are doing a pretty good job, right? Well, it is important to compare our accuracy to what accuracy we would expect to get by chance alone, if predictions were made by a random process rather than using a test's scores. Our selection ratio was 25.4%. How accurate would we be if we randomly selected 25.4% of people to reject? To determine what accuracy we could get by chance alone given the selection ratio and the base rate, we can calculate the chance probability of true positives and the chance probability of true negatives. The probability of a given cell in the confusion matrix is a joint probability—the probability of two events occurring simultaneously. To calculate a joint probability, we multiply the probability of each event.

So, to get the chance expectancies of true positives, we would multiply the respective marginal probabilities, as in Equation 14.1:

$$\begin{aligned} P(TP) &= P(\text{Poor adjustment}) \times P(\text{Reject}) \\ &= BR \times SR \\ &= .05 \times .254 \\ &= .0127 \end{aligned} \tag{14.1}$$

To get the chance expectancies of true negatives, we would multiply the respective marginal probabilities, as in Equation 14.2:

$$\begin{aligned} P(TN) &= P(\text{Good adjustment}) \times P(\text{Retain}) \\ &= (1 - BR) \times (1 - SR) \\ &= .95 \times .746 \\ &= .7087 \end{aligned} \tag{14.2}$$

To get the percent accuracy by chance, we sum the chance expectancies for the correct predictions (TP and TN): $.0127 + .7087 = .7214$. Thus, the percent accuracy you can get by chance alone is 72%. This is because most of our predictions are to retain people, and the base rate of poor adjustment is quite low (.05). Our measure with 78% accuracy provides only a 6% increment in correct predictions. Thus, you cannot judge how good your judgment or prediction is until you know how you would do by random chance.

The chance expectancies for each cell of the confusion matrix are in INSERT

14.6.4 Predicting from the Base Rate

Now, let us consider how well you would do if you were to predict from the base rate. Predicting from the base rate is also called "betting from the base rate", and it involves setting the selection ratio by taking advantage of the base rate so that you go with the most likely outcome in every prediction. Because the base rate is quite low (.05), we could predict from the base rate by selecting no one to reject (i.e., setting the selection ratio at zero). Our percent accuracy by chance if we predict from the base rate would be calculated by multiplying the marginal probabilities, as we did above, but with a new selection ratio, as in Equation 14.3:

$$\begin{aligned} P(TP) &= P(\text{Poor adjustment}) \times P(\text{Reject}) \\ &= BR \times SR \\ &= .05 \times 0 \\ &= 0 \\ \\ P(TN) &= P(\text{Good adjustment}) \times P(\text{Retain}) \\ &= (1 - BR) \times (1 - SR) \\ &= .95 \times 1 \\ &= .95 \end{aligned} \tag{14.3}$$

We sum the chance expectancies for the correct predictions (TP and TN): $0 + .95 = .95$. Thus, our percent accuracy by predicting from the base rate is 95%. This is damning to our measure because it is a much higher accuracy than the accuracy of our measure. That is, we can be much more accurate than our measure simply by predicting from the base rate and selecting no one to reject.

Going with the most likely outcome in every prediction (predicting from the base rate) can be highly accurate (in terms of percent accuracy) as noted by Meehl & Rosen (1955), especially when the base rate is very low or very high. This should serve as an important reminder that we need to compare the accuracy of our measures to the accuracy by (1) random chance and (2) predicting from the base rate. There are several important implications of the impact of base rates on prediction accuracy. One implication is that using the same test in different settings with different base rates will markedly change the accuracy of the test. Oftentimes, using a test will actually *decrease* the predictive accuracy when the base rate deviates greatly from .50. But percent accuracy is not everything. Percent accuracy treats different kinds of errors as if they are equally important. However, the value we place on different kinds of errors may be different, as described next.

14.6.5 Different Kinds of Errors Have Different Costs

Some errors have a high cost, and some errors have a low cost. Among the four decision outcomes, there are two types of errors: false positives and false negatives. The extent to which false positives and false negatives are costly depends on the prediction problem. So, even though you can often be most accurate by going with the base rate, it may be advantageous to use a screening instrument despite lower overall accuracy because of the huge difference in costs of false positives versus false negatives in some cases.

Consider the example of a screening instrument for HIV. False positives would be cases where we said that someone is at high risk of HIV when they are not, whereas false negatives are cases where we said that someone is not at high risk when they actually are. The costs of false positives include a shortage of blood, some follow-up testing, and potentially some anxiety, but that is about it. The costs of false negatives may be people getting HIV. In this case, the costs of false negatives greatly outweigh the costs of false positives, so we use a screening instrument to try to identify the cases at high risk for HIV because of the important consequences of failing to do so, even though using the screening instrument will lower our overall accuracy level.

Another example is when the Central Intelligence Agency (CIA) used a screen for protective typists during wartime to try to detect spies. False positives would be cases where the CIA believes that a person is a spy when they are not, and the CIA does not hire them. False negatives would be cases where the CIA believes that a person is not a spy when they actually are, and the CIA hires them. In this case, a false positive would be fine, but a false negative would be really bad.

How you weigh the costs of different errors depends considerably on the domain and context. Possible costs of false positives to society include: unnecessary and costly treatment with side effects and sending an innocent person to jail (despite our presumption of innocence in the United States criminal justice system that a person is innocent until proven guilty). Possible costs of false negatives to society include: setting a guilty person free, failing to detect a bomb or tumor, and preventing someone from getting treatment who needs it.

The differential costs of different errors also depend on how much flexibility you have in the selection ratio in being able to set a stringent versus loose selection ratio. Consider if there is a high cost of getting rid of people during the selection process. For example, if you must hire 100 people and only 100 people apply for the position, you cannot lose people, so you need to hire even high-risk people. However, if you do not need to hire many people, then you can hire more conservatively.

Any time the selection ratio differs from the base rate, you will make errors. For example, if you reject 25% of applicants, and the base rate of poor adjust-

ment is 5%, then you are making errors of over-rejecting (false positives). By contrast, if you reject 1% of applicants and the base rate of poor adjustment is 5%, then you are making errors of under-rejecting or over-accepting (false negatives).

A low base rate makes it harder to make predictions, and tends to lead to less accurate predictions. For instance, it is very challenging to predict low base rate behaviors, including suicide (Kessler et al., 2020). For this reason, it is likely much more challenging to predict touchdowns—which happen relatively less often—than it is to predict passing/rushing/receiving yards—which are more frequent and continuously distributed.

[EVALUATE EMPIRICALLY]

14.6.6 Sensitivity, Specificity, PPV, and NPV

As described earlier, percent accuracy is not the only important aspect of accuracy. Percent accuracy can be misleading because it is highly influenced by base rates. You can have a high percent accuracy by predicting from the base rate and saying that no one has the condition (if the base rate is low) or that everyone has the condition (if the base rate is high). Thus, it is also important to consider other aspects of accuracy, including sensitivity (SN), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV). We want our predictions to be sensitive to be able to detect the characteristic but also to be specific so that we classify only people actually with the characteristic as having the characteristic.

Let us return to the confusion matrix in INSERT. If we know the frequency of each of the four predicted-actual combinations of the confusion matrix (TP, TN, FP, FN), we can calculate sensitivity, specificity, PPV, and NPV.

Sensitivity is the proportion of those with the characteristic (TP + FN) that we identified with our measure (TP): $\frac{\text{TP}}{\text{TP+FN}} = \frac{86}{86+14} = .86$. Specificity is the proportion of those who do not have the characteristic (TN + FP) that we correctly classify as not having the characteristic (TN): $\frac{\text{TN}}{\text{TN+FP}} = \frac{1{,}478}{1{,}478+422} = .78$. PPV is the proportion of those who we classify as having the characteristic (TP + FP) who actually have the characteristic (TP): $\frac{\text{TP}}{\text{TP+FP}} = \frac{86}{86+422} = .17$. NPV is the proportion of those we classify as not having the characteristic (TN + FN) who actually do not have the characteristic (TN): $\frac{\text{TN}}{\text{TN+FN}} = \frac{1{,}478}{1{,}478+14} = .99$.

Sensitivity, specificity, PPV, and NPV are proportions, and their values therefore range from 0 to 1, where higher values reflect greater accuracy. With sensitivity, specificity, PPV, and NPV, we have a good snapshot of how accurate the measure is at a given cutoff. In our case, our measure is good at finding whom to reject (high sensitivity), but it is rejecting too many people who do

not need to be rejected (lower PPV due to many FPs). Most people whom we classify as having the characteristic do not actually have the characteristic. However, the fact that we are over-rejecting could be okay depending on our goals, for instance, if we do not care about over-dropping (i.e., the PPV being low).

14.6.6.1 Some Accuracy Estimates Depend on the Cutoff

Sensitivity, specificity, PPV, and NPV differ based on the cutoff (i.e., threshold) for classification. Consider the following example. Aliens visit Earth, and they develop a test to determine whether a berry is edible or inedible.

Figure 14.1 depicts the distributions of scores by berry type. Note how there are clearly two distinct distributions. However, the distributions overlap to some degree. Thus, any cutoff will have at least some inaccurate classifications. The extent of overlap of the distributions reflects the amount of measurement error of the measure with respect to the characteristic of interest.

```
#No Cutoff
sampleSize <- 1000

edibleScores <- rnorm(sampleSize, 50, 15)
inedibleScores <- rnorm(sampleSize, 100, 15)

edibleData <- data.frame(
  score = c(
    edibleScores,
    inedibleScores),
  type = c(
    rep("edible", sampleSize),
    rep("inedible", sampleSize)))

cutoff <- 75

hist_edible <- density(
  edibleScores,
  from = 0,
  to = 150) %$% # exposition pipe magrittr::`%$%`
  data.frame(
    x = x,
    y = y) %>%
  mutate(area = x >= cutoff)

hist_edible$type[hist_edible$area == TRUE] <- "edible_FP"
hist_edible$type[hist_edible$area == FALSE] <- "edible_TN"
```

```
hist_inedible <- density(
  inedibleScores,
  from = 0,
  to = 150) %$% # exposition pipe magrittr::`%$%`
  data.frame(
    x = x,
    y = y) %>%
  mutate(area = x < cutoff)

hist_inedible$type[hist_inedible$area == TRUE] <- "inedible_FN"
hist_inedible$type[hist_inedible$area == FALSE] <- "inedible_TP"

density_data <- bind_rows(
  hist_edible,
  hist_inedible)

density_data$type <- factor(
  density_data$type,
  levels = c(
    "edible_TN",
    "inedible_TP",
    "edible_FP",
    "inedible_FN"))

ggplot(
  data = edibleData,
  aes(
    x = score,
    ymin = 0,
    fill = type)) +
  geom_density(alpha = .5) +
  scale_fill_manual(
    name = "Berry Type",
    values = c(
      viridis::viridis(2)[1],
      viridis::viridis(2)[2])) +
  scale_y_continuous(name = "Frequency") +
  theme_bw() +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())
```

Figure 14.1 Distribution of Test Scores by Berry Type.

Figure 14.2 depicts the distributions of scores by berry type with a cutoff. The red line indicates the cutoff—the level above which berries are classified by the test as inedible. There are errors on each side of the cutoff. Below the cutoff, there are some false negatives (blue): inedible berries that are inaccurately classified as edible. Above the cutoff, there are some false positives (green): edible berries that are inaccurately classified as inedible. Costs of false negatives could include sickness or death from eating the inedible berries. Costs of false positives could include taking longer to find food, finding insufficient food, and starvation.

```
#Standard Cutoff
ggplot(
  data = density_data,
  aes(
    x = x,
    ymin = 0,
    ymax = y,
    fill = type)) +
  geom_ribbon(alpha = 1) +
  scale_fill_manual(
    name = "Berry Type",
    values = c(
      viridis::viridis(4)[4],
      viridis::viridis(4)[1],
```

```
      viridis::viridis(4)[3],
      viridis::viridis(4)[2]),
    breaks = c("edible_TN","inedible_TP","edible_FP","inedible_FN"),
    labels = c("Edible: TN","Inedible: TP","Edible: FP","Inedible: FN")) +
  geom_line(aes(y = y)) +
  geom_vline(
    xintercept = cutoff,
    color = "red",
    linewidth = 2) +
  scale_x_continuous(name = "score") +
  scale_y_continuous(name = "Frequency") +
  theme_bw() +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())
```

Figure 14.2 Classifications Based on a Cutoff. Note that some true negatives and true positives are hidden behind the false positives and false negatives.

Based on our assessment goals, we might use a different selection ratio by changing the cutoff. Figure 14.3 depicts the distributions of scores by berry type when we raise the cutoff. There are now more false negatives (blue) and fewer false positives (green). If we raise the cutoff (to be more conservative), the number of false negatives increases and the number of false positives decreases. Consequently, as the cutoff increases, sensitivity and NPV decrease

(because we have more false negatives), whereas specificity and PPV increase (because we have fewer false positives). A higher cutoff could be optimal if the costs of false positives are considered greater than the costs of false negatives. For instance, if the aliens cannot risk eating the inedible berries because the berries are fatal, and there are sufficient edible berries that can be found to feed the alien colony.

```
#Raise the cutoff
cutoff <- 85

hist_edible <- density(
  edibleScores,
  from = 0,
  to = 150) %$% # exposition pipe magrittr::`%$%`
  data.frame(
    x = x,
    y = y) %>%
  mutate(area = x >= cutoff)

hist_edible$type[hist_edible$area == TRUE] <- "edible_FP"
hist_edible$type[hist_edible$area == FALSE] <- "edible_TN"

hist_inedible <- density(
  inedibleScores,
  from = 0,
  to = 150) %$% # exposition pipe magrittr::`%$%`
  data.frame(
    x = x,
    y = y) %>%
  mutate(area = x < cutoff)

hist_inedible$type[hist_inedible$area == TRUE] <- "inedible_FN"
hist_inedible$type[hist_inedible$area == FALSE] <- "inedible_TP"

density_data <- bind_rows(
  hist_edible,
  hist_inedible)

density_data$type <- factor(
  density_data$type,
  levels = c(
    "edible_TN",
    "inedible_TP",
    "edible_FP",
    "inedible_FN"))
```

```
ggplot(
  data = density_data,
  aes(
    x = x,
    ymin = 0,
    ymax = y,
    fill = type)) +
  geom_ribbon(alpha = 1) +
  scale_fill_manual(
    name = "Berry Type",
    values = c(
      viridis::viridis(4)[4],
      viridis::viridis(4)[1],
      viridis::viridis(4)[3],
      viridis::viridis(4)[2]),
    breaks = c("edible_TN","inedible_TP","edible_FP","inedible_FN"),
    labels = c("Edible: TN","Inedible: TP","Edible: FP","Inedible: FN")) +
  geom_line(aes(y = y)) +
  geom_vline(
    xintercept = cutoff,
    color = "red",
    linewidth = 2) +
  scale_x_continuous(name = "score") +
  scale_y_continuous(name = "Frequency") +
  theme_bw() +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())
```

Figure 14.3 Classifications Based on Raising the Cutoff. Note that some true negatives and true positives are hidden behind the false positives and false negatives.

Figure 14.4 depicts the distributions of scores by berry type when we lower the cutoff. There are now fewer false negatives (blue) and more false positives (green). If we lower the cutoff (to be more liberal), the number of false negatives decreases and the number of false positives increases. Consequently, as the cutoff decreases, sensitivity and NPV increase (because we have fewer false negatives), whereas specificity and PPV decrease (because we have more false positives). A lower cutoff could be optimal if the costs of false negatives are considered greater than the costs of false positives. For instance, if the aliens cannot risk missing edible berries because they are in short supply relative to the size of the alien colony, and eating the inedible berries would, at worst, lead to minor, temporary discomfort.

```
#Lower the cutoff
cutoff <- 65

hist_edible <- density(
  edibleScores,
  from = 0,
  to = 150) %$% # exposition pipe magrittr::`%$%`
  data.frame(
    x = x,
    y = y) %>%
```

```
  mutate(area = x >= cutoff)

hist_edible$type[hist_edible$area == TRUE] <- "edible_FP"
hist_edible$type[hist_edible$area == FALSE] <- "edible_TN"

hist_inedible <- density(
  inedibleScores,
  from = 0,
  to = 150) %$% # exposition pipe magrittr::`%$%`
  data.frame(
    x = x,
    y = y) %>%
  mutate(area = x < cutoff)

hist_inedible$type[hist_inedible$area == TRUE] <- "inedible_FN"
hist_inedible$type[hist_inedible$area == FALSE] <- "inedible_TP"

density_data <- bind_rows(
  hist_edible,
  hist_inedible)

density_data$type <- factor(
  density_data$type,
  levels = c(
    "edible_TN",
    "inedible_TP",
    "edible_FP",
    "inedible_FN"))

ggplot(
  data = density_data,
  aes(
    x = x,
    ymin = 0,
    ymax = y,
    fill = type)) +
  geom_ribbon(alpha = 1) +
  scale_fill_manual(
    name = "Berry Type",
    values = c(
      viridis::viridis(4)[4],
      viridis::viridis(4)[1],
      viridis::viridis(4)[3],
      viridis::viridis(4)[2]),
```

```
    breaks = c("edible_TN","inedible_TP","edible_FP","inedible_FN"),
    labels = c("Edible: TN","Inedible: TP","Edible: FP","Inedible: FN")) +
  geom_line(aes(y = y)) +
  geom_vline(
    xintercept = cutoff,
    color = "red",
    linewidth = 2) +
  scale_x_continuous(name = "score") +
  scale_y_continuous(name = "Frequency") +
  theme_bw() +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank())
```

Figure 14.4 Classifications Based on Lowering the Cutoff. Note that some true negatives and true positives are hidden behind the false positives and false negatives.

In sum, sensitivity and specificity differ based on the cutoff for classification. If we raise the cutoff, sensitivity and PPV increase (due to fewer false positives), whereas sensitivity and NPV decrease (due to more false negatives). If we lower the cutoff, sensitivity and NPV increase (due to fewer false negatives), whereas specificity and PPV decrease (due to more false positives). Thus, the optimal cutoff depends on how costly each type of error is: false negatives and false positives. If false negatives are more costly than false positives, we

would set a low cutoff. If false positives are more costly than false negatives, we would set a high cutoff.

14.6.7 Signal Detection Theory

Signal detection theory (SDT) is a probability-based theory for the detection of a given stimulus (signal) from a stimulus set that includes non-target stimuli (noise). SDT arose through the development of radar (**RA**dio **D**etection **A**nd **R**anging) and sonar (**SO**und **N**avigation **A**nd **R**anging) in World War II based on research on sensory-perception research. The military wanted to determine which objects on radar/sonar were enemy aircraft/submarines, and which were noise (e.g., different object in the environment or even just the weather itself). SDT allowed determining how many errors operators made (how accurate they were) and decomposing errors into different kinds of errors. SDT distinguishes between sensitivity and bias. In SDT, *sensitivity* (or discriminability) is how well an assessment distinguishes between a target stimulus and non-target stimuli (i.e., how well the assessment detects the target stimulus amid non-target stimuli). *Bias* is the extent to which the probability of a selection decision from the assessment is higher or lower than the true rate of the target stimulus.

Some radar/sonar operators were not as sensitive to the differences between signal and noise, due to factors such as age, ability to distinguish gradations of a signal, etc. People who showed low sensitivity (i.e., who were not as successful at distinguishing between signal and noise) were screened out because the military perceived sensitivity as a skill that was not easily taught. By contrast, other operators could distinguish signal from noise, but their threshold was too low or high—they could take in information, but their decisions tended to be wrong due to systematic bias or poor calibration. That is, they systematically over-rejected or under-rejected stimuli. Over-rejecting leads to many false negatives (i.e., saying that a stimulus is safe when it is not). Under-rejecting leads to many false positives (i.e., saying that a stimulus is harmful when it is not). A person who showed good sensitivity but systematic bias was considered more teach-able than a person who showed low sensitivity. Thus, radar and sonar operators were selected based on their sensitivity to distinguish signal from noise, and then were trained to improve the calibration so they reduce their systematic bias and do not systematically over- or under-reject.

Although SDT was originally developed for use in World War II, it now plays an important role in many areas of science and medicine. A medical application of SDT is tumor detection in radiology. Another application of SDT in society is using x-ray to detect bombs or other weapons. An example of applying SDT to fantasy football could be in the prediction (and evaluation) of whether or not a player scores a touchdown in a game.

SDT metrics of sensitivity include d' ("d-prime"), A (or A'), and the area under the receiver operating characteristic (ROC) curve. SDT metrics of bias include β (beta), c, and b.

14.6.7.1 Receiver Operating Characteristic (ROC) Curve

The x-axis of the ROC curve is the false alarm rate or false positive rate (1− specificity). The y-axis is the hit rate or true positive rate (sensitivity). We can trace the ROC curve as the combination between sensitivity and specificity at every possible cutoff. At a cutoff of zero (top right of ROC curve), we calculate sensitivity (1.0) and specificity (0) and plot it. At a cutoff of zero, the assessment tells us to make an action for every stimulus (i.e., it is the most liberal). We then gradually increase the cutoff, and plot sensitivity and specificity at each cutoff. As the cutoff increases, sensitivity decreases and specificity increases. We end at the highest possible cutoff, where the sensitivity is 0 and the specificity is 1.0 (i.e., we never make an action; i.e., it is the most conservative). Each point on the ROC curve corresponds to a pair of hit and false alarm rates (sensitivity and specificity) resulting from a specific cutoff value. Then, we can draw lines or a curve to connect the points.

INSERT depicts an empirical ROC plot where lines are drawn to connect the hit and false alarm rates.

INSERT depicts an ROC curve where a smoothed and fitted curve is drawn to connect the hit and false alarm rates.

14.6.7.1.1 Area Under the ROC Curve

ROC methods can be used to compare and compute the discriminative power of measurement devices free from the influence of selection ratios, base rates, and costs and benefits. An ROC analysis yields a quantitative index of how well an index predicts a signal of interest or can discriminate between different signals. ROC analysis can help tell us how often our assessment would be correct. If we randomly pick two observations, and we were right once and wrong once, we were 50% accurate. But this would be a useless measure because it reflects chance responding.

The geometrical area under the ROC curve reflects the discriminative accuracy of the measure. The index is called the **a**rea **u**nder the **c**urve (AUC) of an ROC curve. AUC quantifies the discriminative power of an assessment. AUC is the probability that a randomly selected target and a randomly selected non-target is ranked correctly by the assessment method. AUC values range from 0.0 to 1.0, where chance accuracy is 0.5 as indicated by diagonal line in the ROC curve. That is, a measure can be useful to the extent that its ROC curve is above the diagonal line (i.e., its discriminative accuracy is above chance).

AUC is a threshold-independent accuracy index that applies across all possible cutoff values.

Figure 14.5 depicts ROC curves with a range of AUC values.

```
set.seed(52242)

auc60 <- petersenlab::simulateAUC(.60, 50000)
auc70 <- petersenlab::simulateAUC(.70, 50000)
auc80 <- petersenlab::simulateAUC(.80, 50000)
auc90 <- petersenlab::simulateAUC(.90, 50000)
auc95 <- petersenlab::simulateAUC(.95, 50000)
auc99 <- petersenlab::simulateAUC(.99, 50000)

plot(
  pROC::roc(
    y ~ x,
    auc60,
    smooth = TRUE),
  legacy.axes = TRUE,
  print.auc = TRUE,
  print.auc.x = .52,
  print.auc.y = .61,
  print.auc.pattern = "%.2f")

plot(
  pROC::roc(
    y ~ x,
    auc70,
    smooth = TRUE),
  legacy.axes = TRUE,
  print.auc = TRUE,
  print.auc.x = .6,
  print.auc.y = .67,
  print.auc.pattern = "%.2f",
  add = TRUE)

plot(
  pROC::roc(
    y ~ x,
    auc80,
    smooth = TRUE),
  legacy.axes = TRUE,
  print.auc = TRUE,
  print.auc.x = .695,
  print.auc.y = .735,
```

```
  print.auc.pattern = "%.2f",
  add = TRUE)

plot(
  pROC::roc(
    y ~ x,
    auc90,
    smooth = TRUE),
  legacy.axes = TRUE,
  print.auc = TRUE,
  print.auc.x = .805,
  print.auc.y = .815,
  print.auc.pattern = "%.2f",
  add = TRUE)

plot(
  pROC::roc(
    y ~ x,
    auc95,
    smooth = TRUE),
  legacy.axes = TRUE,
  print.auc = TRUE,
  print.auc.x = .875,
  print.auc.y = .865,
  print.auc.pattern = "%.2f",
  add = TRUE)

plot(
  pROC::roc(
    y ~ x,
    auc99,
    smooth = TRUE),
  legacy.axes = TRUE,
  print.auc = TRUE,
  print.auc.x = .94,
  print.auc.y = .94,
  print.auc.pattern = "%.2f",
  add = TRUE)
```

Figure 14.5 Receiver Operating Characteristic (ROC) Curves for Various Levels of Area Under The ROC Curve (AUC) for Various Measures.

As an example, given an AUC of .75, this says that the overall score of an individual who has the characteristic in question will be higher 75% of the time than the overall score of an individual who does not have the characteristic. In lay terms, AUC provides the probability that we will classify correctly based on our instrument if we were to randomly pick one good and one bad outcome. AUC is a stronger index of accuracy than percent accuracy, because you can have high percent accuracy just by going with the base rate. AUC tells us how much better than chance a measure is at discriminating outcomes. AUC is useful as a measure of general discriminative accuracy, and it tells us how accurate a measure is at all possible cutoffs. Knowing the accuracy of a measure at all possible cutoffs can be helpful for selecting the optimal cutoff, given the goals of the assessment. In reality, however, we may not be interested in all cutoffs because not all errors are equal in their costs.

If we lower the base rate, we would need a larger sample to get enough people to classify into each group. SDT/ROC methods are traditionally about dichotomous decisions (yes/no), not graded judgments. SDT/ROC methods can get messy with ordinal data that are more graded because you would have an AUC curve for each ordinal grouping.

14.6.8 Accuracy Indices

There are various accuracy indices we can use to evaluate the accuracy of predictions for categorical outcome variables. We have already described several accuracy indices, including percent accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and area under the ROC curve. We describe these and other indices in greater detail below.

The `petersenlab`[1] package (Petersen, 2024a) contains the `accuracyAtCutoff()` function that computes many accuracy indices for the prediction of categorical outcome variables.

```
#petersenlab::accuracyAtCutoff()
```

The `petersenlab`[2] package (Petersen, 2024a) contains the `accuracyAtEachCutoff()` function that computes many accuracy indices for the prediction of categorical outcome variables at each possible cutoff.

```
#petersenlab::accuracyAtEachCutoff()
```

There are also test calculators available online:

- http://araw.mede.uic.edu/cgi-bin/testcalc.pl
- https://dlrs.shinyapps.io/shinyDLRs

14.6.8.1 Confusion Matrix aka 2x2 Accuracy Table aka Cross-Tabulation aka Contingency Table

A confusion matrix (aka 2x2 accuracy table, cross-tabulation table, or contigency table) is a matrix for categorical data that presents the predicted outcome on one dimension and the actual outcome (truth) on the other dimension. If the predictions and outcomes are dichotomous, the confusion matrix is a 2x2 matrix with two rows and two columns that represent four possible predicted-actual combinations (decision outcomes). In such a case, the confusion matrix provides a tabular count of each type of accurate cases (true

[1] https://github.com/DevPsyLab/petersenlab
[2] https://github.com/DevPsyLab/petersenlab

positives and true negatives) versus the number of each type of error (false positives and false negatives), as shown in INSERT. An example of a confusion matrix is in INSERT.

14.6.8.1.1 Number

```
#table(mydata$diagnosisFactor, mydata$diseaseFactor)
```

14.6.8.1.2 Number with margins added

```
#addmargins(table(mydata$diagnosisFactor, mydata$diseaseFactor))
```

14.6.8.1.3 Proportions

```
#prop.table(table(mydata$diagnosisFactor, mydata$diseaseFactor))
```

14.6.8.1.4 Proportions with margins added

```
#addmargins(prop.table(table(mydata$diagnosisFactor, mydata$diseaseFactor)))
```

14.6.8.2 True Positives (TP)

True positives (TPs) are instances in which a positive classification (e.g., stating that a disease is present for a person) is correct—that is, the test says that a classification is present, and the classification is present. True positives are also called valid positives (VPs) or hits. Higher values reflect greater accuracy. The formula for true positives is in Equation 14.4:

$$\text{TP} = \text{BR} \times \text{SR} \times N \tag{14.4}$$

14.6.8.3 True Negatives (TN)

True negatives (TNs) are instances in which a negative classification (e.g., stating that a disease is absent for a person) is correct—that is, the test

says that a classification is not present, and the classification is actually not present. True negatives are also called valid negatives (VNs) or correct rejections. Higher values reflect greater accuracy. The formula for true negatives is in Equation 14.5:

$$\mathrm{TN} = (1 - \mathrm{BR}) \times (1 - \mathrm{SR}) \times N \tag{14.5}$$

14.6.8.4 False Positives (FP)

False positives (FPs) are instances in which a positive classification (e.g., stating that a disease is present for a person) is incorrect—that is, the test says that a classification is present, and the classification is not present. False positives are also called false alarms (FAs). Lower values reflect greater accuracy. The formula for false positives is in Equation Equation 14.6:

$$\mathrm{FP} = (1 - \mathrm{BR}) \times \mathrm{SR} \times N \tag{14.6}$$

14.6.8.5 False Negatives (FN)

False negatives (FNs) are instances in which a negative classification (e.g., stating that a disease is absent for a person) is incorrect—that is, the test says that a classification is not present, and the classification is present. False negatives are also called misses. Lower values reflect greater accuracy. The formula for false negatives is in Equation 14.7:

$$\mathrm{FN} = \mathrm{BR} \times (1 - \mathrm{SR}) \times N \tag{14.7}$$

14.6.8.6 Selection Ratio (SR)

The selection ratio (SR) is the marginal probability of selection, independent of other things: $P(R_i)$. It is not an index of accuracy, per se. In medicine, the selection ratio is the proportion of people who test positive for the disease. In fantasy football, the selection ratio is the proportion of players who you predict will show a given outcome. For example, if you are trying to predict the players who will score a touchdown in a game, the selection ratio is the proportion of players who you predict will score a touchdown. The formula for calculating the selection ratio is in Equation 14.8.

$$\begin{aligned} \mathrm{SR} &= P(R_i) \\ &= \frac{\mathrm{TP} + \mathrm{FP}}{N} \end{aligned} \tag{14.8}$$

14.6.8.7 Base Rate (BR)

The base rate (BR) of a classification is its marginal probability, independent of other things: $P(C_i)$. It is not an index of accuracy, per se. In medicine, the base rate of a disease is its prevalence in the population, as in Equation 14.9. Without additional information, the base rate is used as the initial *pretest probability*. In fantasy football, the base rate is the proportion of players who actually show the particular outcome. For example, if you are trying to predict the players who will score a touchdown in a game, the base rate is the proportion of players who actually score a touchdown in the game. The formula for calculating the selection ratio is in Equation 14.9.

$$\begin{aligned} \text{BR} &= P(C_i) \\ &= \frac{\text{TP} + \text{FN}}{N} \end{aligned} \tag{14.9}$$

14.6.8.8 Pretest Odds

The pretest odds of a classification can be estimated using the pretest probability (i.e., base rate). To convert a probability to odds, divide the probability by one minus that probability, as in Equation 14.10.

$$\text{pretest odds} = \frac{\text{pretest probability}}{1 - \text{pretest probability}} \tag{14.10}$$

14.6.8.9 Percent Accuracy

Percent Accuracy is also called overall accuracy. Higher values reflect greater accuracy. The formula for percent accuracy is in Equation 14.11. Percent accuracy has several problems. First, it treats all errors (FP and FN) as equally important. However, in practice, it is rarely the case that false positives and false negatives are equally important. Second, percent accuracy can be misleading because it is highly influenced by base rates. You can have a high percent accuracy by predicting from the base rate and saying that no one has the characteristic (if the base rate is low) or that everyone has the characteristic (if the base rate is high). Thus, it is also important to consider other aspects of accuracy.

$$\text{Percent Accuracy} = 100\% \times \frac{\text{TP} + \text{TN}}{N} \tag{14.11}$$

14.6.8.10 Percent Accuracy by Chance

The formula for calculating percent accuracy by chance is in Equation 14.12.

$$\begin{aligned}\text{Percent Accuracy by Chance} &= 100\% \times [P(\text{TP}) + P(\text{TN})] \\ &= 100\% \times \{(\text{BR} \times \text{SR}) + [(1 - \text{BR}) \times (1 - \text{SR})]\}\end{aligned} \tag{14.12}$$

14.6.8.11 Percent Accuracy Predicting from the Base Rate

Predicting from the base rate is going with the most likely outcome in every prediction. If the base rate is less than .50, it would involve predicting that the condition is absent for every case. If the base rate is .50 or above, it would involve predicting that the condition is present for every case. Predicting from the base rate is a special case of percent accuracy by chance when the selection ratio is set to either one (if the base rate $\geq .5$) or zero (if the base rate $< .5$).

14.6.8.12 Relative Improvement Over Chance (RIOC)

Relative improvement over chance (RIOC) is a prediction's improvement over chance as a proportion of the maximum possible improvement over chance, as described by Farrington & Loeber (1989). Higher values reflect greater accuracy. The formula for calculating RIOC is in Equation 14.13.

$$\text{relative improvement over chance (RIOC)} = \frac{\text{total correct} - \text{chance correct}}{\text{maximum correct} - \text{chance correct}} \tag{14.13}$$

14.6.8.13 Relative Improvement Over Predicting from the Base Rate

Relative improvement over predicting from the base rate is a prediction's improvement over predicting from the base rate as a proportion of the maximum possible improvement over predicting from the base rate. Higher values reflect greater accuracy. The formula for calculating relative improvement over predicting from the base rate is in Equation 14.14.

$$\text{relative improvement over predicting from base rate} = \frac{\text{total correct} - \text{correct by predicting from base rate}}{\text{maximum correct} - \text{correct by predicting from base rate}} \tag{14.14}$$

14.6.8.14 Sensitivity (SN)

Sensitivity (SN) is also called true positive rate (TPR), hit rate (HR), or recall. Sensitivity is the conditional probability of a positive test given that the person has the condition: $P(R|C)$. Higher values reflect greater accuracy. The formula for calculating sensitivity is in Equation 14.15. As described in

Section Section 14.6.6.1, as the cutoff increases (becomes more conservative), sensitivity decreases. As the cutoff decreases, sensitivity increases.

$$\begin{aligned}\text{sensitivity (SN)} &= P(R|C) \\ &= \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{\text{TP}}{N \times \text{BR}} = 1 - \text{FNR}\end{aligned} \tag{14.15}$$

14.6.8.15 Specificity (SP)

Specificity (SP) is also called true negative rate (TNR) or selectivity. Specificity is the conditional probability of a negative test given that the person does not have the condition: $P(\text{not } R|\text{not } C)$. Higher values reflect greater accuracy. The formula for calculating specificity is in Equation 14.16. As described in Section Section 14.6.6.1, as the cutoff increases (becomes more conservative), specificity increases. As the cutoff decreases, specificity decreases.

$$\begin{aligned}\text{specificity (SP)} &= P(\text{not } R|\text{not } C) \\ &= \frac{\text{TN}}{\text{TN}+\text{FP}} = \frac{\text{TN}}{N(1-\text{BR})} = 1 - \text{FPR}\end{aligned} \tag{14.16}$$

14.6.8.16 False Negative Rate (FNR)

The false negative rate (FNR) is also called the miss rate. The false negative rate is the conditional probability of a negative test given that the person has the condition: $P(\text{not } R|C)$. Lower values reflect greater accuracy. The formula for calculating false negative rate is in Equation 14.17.

$$\begin{aligned}\text{false negative rate (FNR)} &= P(\text{not } R|C) \\ &= \frac{\text{FN}}{\text{FN}+\text{TP}} = \frac{\text{FN}}{N \times \text{BR}} = 1 - \text{TPR}\end{aligned} \tag{14.17}$$

14.6.8.17 False Positive Rate (FPR)

The false positive rate (FPR) is also called the false alarm rate (FAR) or fall-out. The false positive rate is the conditional probability of a positive test given that the person does not have the condition: $P(R|\text{not } C)$. Lower values reflect greater accuracy. The formula for calculating false positive rate is in Equation 14.18:

$$\begin{aligned}\text{false positive rate (FPR)} &= P(R|\text{not } C) \\ &= \frac{\text{FP}}{\text{FP}+\text{TN}} = \frac{\text{FP}}{N(1-\text{BR})} = 1 - \text{TNR}\end{aligned} \tag{14.18}$$

14.6.8.18 Positive Predictive Value (PPV)

The positive predictive value (PPV) is also called the positive predictive power (PPP) or precision. Many people confuse sensitivity ($P(R|C)$) with its inverse conditional probability, PPV ($P(C|R)$). PPV is the conditional probability of having the condition given a positive test: $P(C|R)$. Higher values reflect greater accuracy. The formula for calculating positive predictive value is in Equation 14.19.

PPV can be low even when sensitivity is high because it depends not only on sensitivity, but also on specificity and the base rate. Because PPV depends on the base rate, PPV is not an intrinsic property of a measure. The same measure will have a different PPV in different contexts with different base rates (Treat & Viken, 2023). As described in Section Section 14.6.6.1, as the base rate increases, PPV increases. As the base rate decreases, PPV decreases. PPV also differs as a function of the cutoff. As described in Section Section 14.6.6.1, as the cutoff increases (becomes more conservative), PPV increases. As the cutoff decreases (becomes more liberal), PPV decreases.

$$\begin{aligned}\text{positive predictive value (PPV)} &= P(C|R)\\ &= \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{\text{TP}}{N \times \text{SR}}\\ &= \frac{\text{sensitivity} \times \text{BR}}{\text{sensitivity} \times \text{BR} + [(1-\text{specificity}) \times (1-\text{BR})]}\end{aligned} \tag{14.19}$$

14.6.8.19 Negative Predictive Value (NPV)

The negative predictive value (NPV) is also called the negative predictive power (NPP). Many people confuse specificity ($P(\text{not } R|\text{not } C)$) with its inverse conditional probability, NPV ($P(\text{not } C|\text{not } R)$). NPV is the conditional probability of not having the condition given a negative test: $P(\text{not } C|\text{not } R)$. Higher values reflect greater accuracy. The formula for calculating negative predictive value is in Equation 14.20.

NPV can be low even when specificity is high because it depends not only on specificity, but also on sensitivity and the base rate. Because NPV depends on the base rate, NPV is not an intrinsic property of a measure. The same measure will have a different NPV in different contexts with different base rates (Treat & Viken, 2023). As described in Section Section 14.6.6.1, as the base rate increases, NPV decreases. As the base rate decreases, NPV increases. NPV also differs as a function of the cutoff. As described in Section Section 14.6.6.1, as the cutoff increases (becomes more conservative), NPV decreases. As the cutoff decreases (becomes more liberal), NPV decreases.

$$\begin{aligned}\text{negative predictive value (NPV)} &= P(\text{not } C|\text{not } R)\\ &= \frac{\text{TN}}{\text{TN}+\text{FN}} = \frac{\text{TN}}{N(1\text{ - SR})}\\ &= \frac{\text{specificity} \times (1-\text{BR})}{\text{specificity} \times (1-\text{BR}) + [(1-\text{sensitivity}) \times \text{BR})]}\end{aligned} \tag{14.20}$$

14.6.8.20 False Discovery Rate (FDR)

Many people confuse the false positive rate ($P(R|\text{not } C)$) with its inverse conditional probability, the false discovery rate ($P(\text{not } C|R)$). The false discovery rate (FDR) is the conditional probability of not having the condition given a positive test: $P(\text{not } C|R)$. Lower values reflect greater accuracy. The formula for calculating false discovery rate is in Equation 14.21.

$$\begin{aligned}\text{false discovery rate (FDR)} &= P(\text{not } C|R)\\ &= \frac{\text{FP}}{\text{FP}+\text{TP}} = 1-\text{PPV}\end{aligned} \tag{14.21}$$

14.6.8.21 False Omission Rate (FOR)

Many people confuse the false negative rate ($P(\text{not } R|C)$) with its inverse conditional probability, the false omission rate ($P(C|\text{not } R)$). The false omission rate (FOR) is the conditional probability of having the condition given a negative test: $P(C|\text{not } R)$. Lower values reflect greater accuracy. The formula for calculating false omission rate is in Section 14.6.8.21.

$$\begin{aligned}\text{false omission rate (FOR)} &= P(C|\text{not } R)\\ &= \frac{\text{FN}}{\text{FN}+\text{TN}} = 1-\text{NPV}\end{aligned}$$

{#sec-falseOmissionRate}

14.6.8.22 Youden's J Statistic

Youden's J statistic is also called Youden's Index or informedness. Youden's J statistic is the sum of sensitivity and specificity (and subtracting one). Higher values reflect greater accuracy. The formula for calculating Youden's J statistic is in Equation 14.22.

$$\text{Youden's J statistic} = \text{sensitivity} + \text{specificity} - 1 \tag{14.22}$$

14.6.8.23 Balanced Accuracy

Balanced accuracy is the average of sensitivity and specificity. Higher values reflect greater accuracy. The formula for calculating balanced accuracy is in Equation 14.23.

$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2} \tag{14.23}$$

14.6.8.24 F-Score

The F-score combines precision (positive predictive value) and recall (sensitivity), where β indicates how many times more important sensitivity is than the positive predictive value. If sensitivity and the positive predictive value are equally important, $\beta = 1$, and the F-score is called the F_1 score. Higher values reflect greater accuracy. The formula for calculating the F-score is in Equation 14.24.

$$\begin{aligned} F_\beta &= (1+\beta^2) \cdot \frac{\text{positive predictive value} \cdot \text{sensitivity}}{(\beta^2 \cdot \text{positive predictive value}) + \text{sensitivity}} \\ &= \frac{(1+\beta^2) \cdot \text{TP}}{(1+\beta^2) \cdot \text{TP} + \beta^2 \cdot \text{FN} + \text{FP}} \end{aligned} \tag{14.24}$$

The formula for calculating the F_1 score is in Equation 14.25.

$$\begin{aligned} F_1 &= \frac{2 \cdot \text{positive predictive value} \cdot \text{sensitivity}}{(\text{positive predictive value}) + \text{sensitivity}} \\ &= \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}} \end{aligned} \tag{14.25}$$

14.6.8.25 Matthews Correlation Coefficient (MCC)

The Matthews correlation coefficient (MCC) is also called the phi coefficient. It is a correlation coefficient between predicted and observed values from a binary classification. Higher values reflect greater accuracy. The formula for calculating the MCC is in Equation 14.26.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{14.26}$$

14.6.8.26 Diagnostic Odds Ratio

The diagnostic odds ratio is the odds of a positive test among people with the condition relative to the odds of a positive test among people without the

condition. Higher values reflect greater accuracy. The formula for calculating the diagnostic odds ratio is in Equation 14.27. If the predictor is bad, the diagnostic odds ratio could be less than one, and values can go up from there. If the diagnostic odds ratio is greater than 2, we take the odds ratio seriously because we are twice as likely to predict accurately than inaccurately. However, the diagnostic odds ratio ignores/hides base rates. When interpreting the diagnostic odds ratio, it is important to keep in mind the practical significance, because otherwise it is not very meaningful. Consider a risk factor that has a diagnostic odds ratio of 3 for tuberculosis, i.e., it puts you at 3 times as likely to develop tuberculosis. The prevalence of tuberculosis is relatively low. Assuming the prevalence of tuberculosis is less than 1/10th of 1%, your risk of developing tuberculosis is still very low even if the risk factor (with a diagnostic odds ratio of 3) is present.

$$\begin{aligned}\text{diagnostic odds ratio} &= \frac{\text{TP} \times \text{TN}}{\text{FP} \times \text{FN}} \\ &= \frac{\text{sensitivity} \times \text{specificity}}{(1 - \text{sensitivity}) \times (1 - \text{specificity})} \\ &= \frac{\text{PPV} \times \text{NPV}}{(1 - \text{PPV}) \times (1 - \text{NPV})} \\ &= \frac{\text{LR+}}{\text{LR}-}\end{aligned} \tag{14.27}$$

14.6.8.27 Diagnostic Likelihood Ratio

The diagnostic likelihood ratio is described in Section 13.5.2.1. There are two types of diagnostic likelihood ratios: the positive likelihood ratio and the negative likelihood ratio.

14.6.8.27.1 Positive Likelihood Ratio (LR+)

The positive likelihood ratio (LR+) is described in Section 13.5.2.1.1. The formula for calculating the positive likelihood ratio is in Equation 13.14.

14.6.8.27.2 Negative Likelihood Ratio (LR−)

The negative likelihood ratio (LR−) is described in Section 13.5.2.1.2. The formula for calculating the negative likelihood ratio is in Equation 13.14.

14.6.8.28 Posttest Odds

As presented in Equation 13.13, the posttest (or posterior) odds are equal to the pretest odds multiplied by the likelihood ratio. The posttest odds and posttest probability can be useful to calculate when the pretest probability is different from the pretest probability (or prevalence) of the classification. For instance, you might use a different pretest probability if a test result is already known and you want to know the updated posttest probability after conducting a second test. The formula for calculating posttest odds is in Equation 14.28.

$$\text{posttest odds} = \text{pretest odds} \times \text{likelihood ratio} \tag{14.28}$$

For calculating the posttest odds of a true positive compared to a false positive, we use the positive likelihood ratio below. We would use the negative likelihood ratio if we wanted to calculate the posttest odds of a false negative compared to a true negative.

14.6.8.29 Posttest Probability

The posttest probability is the probability of having the characteristic given a test result. When the base rate is used as the pretest probability, the posttest probability given a positive test is equal to positive predictive value. To convert odds to a probability, divide the odds by one plus the odds, as is in Equation 14.29.

$$\text{posttest probability} = \frac{\text{posttest odds}}{1 + \text{posttest odds}} \tag{14.29}$$

14.6.8.30 Mean Difference Between Predicted and Observed Values

The mean difference between predicted values versus observed values at a given cutoff is an index of miscalibration of predictions at that cutoff. It is called "calibration-in-the-small" (as opposed to calibration-in-the-large, which spans all cutoffs). Values closer to zero reflect greater accuracy. Values above zero indicate that the predicted values are, on average, greater than the observed values. Values below zero indicate that the observed values are, on average, greater than the predicted values.

14.7 Threshold-Independent Accuracy Indices

This section describes threshold-independent indexes of accuracy. That is, each index of accuracy described in this section provides a single numerical index of accuracy that aggregates the accuracy across all possible cutoffs. The `petersenlab`[3] package (Petersen, 2024a) contains the `accuracyOverall()` function that computes many threshold-independent accuracy indices.

14.7.1 General Prediction Accuracy

There are many metrics of general prediction accuracy. When thinking about which metric(s) may be best for a given problem, it is important to consider the purpose of the assessment. The estimates of general prediction accuracy are separated below into scale-dependent and scale-independent accuracy estimates.

14.7.1.1 Scale-Dependent Accuracy Estimates

The estimates of prediction accuracy described in this section are scale-dependent. These accuracy estimates depend on the unit of measurement and therefore cannot be compared across measures with different scales or across data sets.

14.7.1.1.1 Mean Error

Here, "error" (e) is the difference between the predicted and observed value for a given individual (i). Mean error (ME; also known as bias) is the mean difference between the predicted and observed values across individuals (i), that is, the mean of the errors across individuals (e_i). Values closer to zero reflect greater accuracy. If mean error is above zero, it indicates that predicted values are, on average, greater than observed values (i.e., overestimating errors). If mean error is below zero, it indicates that predicted values are, on average, less than observed values (i.e., underestimating errors). If both overestimating and under-estimating errors are present, however, they can cancel each other out. As a result, even with a mean error of zero, there can still be considerable error present. Thus, although mean error can be helpful for

[3] https://github.com/DevPsyLab/petersenlab

examining whether predictions systematically under- or over-estimate the actual scores, other forms of accuracy are necessary to examine the *extent* of error. The formula for mean error is in Equation 14.30:

$$\begin{aligned}\text{mean error} &= \frac{\sum_{i=1}^{n}(\text{predicted}_i - \text{observed}_i)}{n} \\ &= \text{mean}(e_i)\end{aligned} \tag{14.30}$$

14.7.1.1.2 Mean Absolute Error (MAE)

Mean absolute error (MAE) is the mean of the absolute value of differences between the predicted and observed values across individuals, that is, the mean of the absolute value of errors. Smaller MAE values (closer to zero) reflect greater accuracy. MAE is preferred over root mean squared error (RMSE) when you want to give equal weight to all errors and when the outliers have considerable impact. The formula for MAE is in Equation 14.31:

$$\begin{aligned}\text{mean absolute error (MAE)} &= \frac{\sum_{i=1}^{n}|\text{predicted}_i - \text{observed}_i|}{n} \\ &= \text{mean}(|e_i|)\end{aligned} \tag{14.31}$$

14.7.1.1.3 Mean Squared Error (MSE)

Mean squared error (MSE) is the mean of the square of the differences between the predicted and observed values across individuals, that is, the mean of the squared value of errors. Smaller MSE values (closer to zero) reflect greater accuracy. MSE penalizes larger errors more heavily than smaller errors (unlike MAE). However, MSE is sensitive to outliers and can be impacted if the errors are skewed. The formula for MSE is in Equation 14.32:

$$\begin{aligned}\text{mean squared error (MSE)} &= \frac{\sum_{i=1}^{n}(\text{predicted}_i - \text{observed}_i)^2}{n} \\ &= \text{mean}(e_i^2)\end{aligned} \tag{14.32}$$

14.7.1.1.4 Root Mean Squared Error (RMSE)

Root mean squared error (RMSE) is the square root of the mean of the square of the differences between the predicted and observed values across individuals, that is, the root mean squared value of errors. Smaller RMSE values (closer to zero) reflect greater accuracy. RMSE penalizes larger errors more heavily than smaller errors (unlike MAE). However, RMSE is sensitive to outliers and can be impacted if the errors are skewed. The formula for RMSE is in Equation 14.33:

$$\begin{aligned}\text{root mean squared error (RMSE)} &= \sqrt{\frac{\sum_{i=1}^{n}(\text{predicted}_i - \text{observed}_i)^2}{n}} \\ &= \sqrt{\text{mean}(e_i^2)}\end{aligned} \tag{14.33}$$

14.7.1.2 Scale-Independent Accuracy Estimates

The estimates of prediction accuracy described in this section are intended to be scale-*independent* (unit-free) so the accuracy estimates can be compared across measures with different scales or across data sets (Hyndman & Athanasopoulos, 2021).

14.7.1.2.1 Mean Percentage Error (MPE)

Mean percentage error (MPE) values closer to zero reflect greater accuracy. The formula for percentage error is in Equation 14.34:

$$\text{percentage error } (p_i) = \frac{100\% \times (\text{observed}_i - \text{predicted}_i)}{\text{observed}_i} \tag{14.34}$$

We then take the mean of the percentage errors to get MPE. The formula for MPE is in Equation 14.35:

$$\begin{aligned}\text{mean percentage error (MPE)} &= \frac{100\%}{n}\sum_{i=1}^{n}\frac{\text{observed}_i - \text{predicted}_i}{\text{observed}_i} \\ &= \text{mean(percentage error)} \\ &= \text{mean}(p_i)\end{aligned} \tag{14.35}$$

Note: MPE is undefined when one or more of the observed values equals zero, due to division by zero. The `accuracyOverall()` function of the `petersenlab`[4]

[4] https://github.com/DevPsyLab/petersenlab

package (Petersen, 2024a) provides the option in the function to drop undefined values so you can still generate an estimate of accuracy despite undefined values.

14.7.1.2.2 Mean Absolute Percentage Error (MAPE)

Smaller mean absolute percentage error (MAPE) values (closer to zero) reflect greater accuracy. The formula for MAPE is in Equation 14.36. MAPE is asymmetric because it overweights underestimates and underweights overestimates. MAPE can be preferable to symmetric mean absolute percentage error (sMAPE) if there are no observed values of zero and if you want to emphasize the importance of underestimates (relative to overestimates).

$$\begin{aligned}\text{mean absolute percentage error (MAPE)} &= \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\text{observed}_i - \text{predicted}_i}{\text{observed}_i}\right| \\ &= \text{mean}(|\text{percentage error}|) \\ &= \text{mean}(|p_i|)\end{aligned} \tag{14.36}$$

Note: MAPE is undefined when one or more of the observed values equals zero, due to division by zero. The `accuracyOverall()` function of the `petersenlab`[5] package (Petersen, 2024a) provides the option in the function to drop undefined values so you can still generate an estimate of accuracy despite undefined values.

14.7.1.2.3 Symmetric Mean Absolute Percentage Error (sMAPE)

Unlike MAPE, symmetric mean absolute percentage error (sMAPE) is symmetric because it equally weights underestimates and overestimates. Smaller sMAPE values (closer to zero) reflect greater accuracy. The formula for sMAPE is in Equation 14.37:

$$\text{symmetric mean absolute percentage error (sMAPE)} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\text{predicted}_i - \text{observed}_i|}{|\text{predicted}_i| + |\text{observed}_i|} \tag{14.37}$$

Note: sMAPE is undefined when one or more of the individuals has a prediction–observed combination such that the sum of the absolute value of the predicted value and the absolute value of the observed value equals zero ($|\text{predicted}_i| + |\text{observed}_i|$), due to division by zero. The `accuracyOverall()`

[5] https://github.com/DevPsyLab/petersenlab

function of the `petersenlab`[6] package (Petersen, 2024a) provides the option in the function to drop undefined values so you can still generate an estimate of accuracy despite undefined values.

14.7.1.2.4 Mean Absolute Scaled Error (MASE)

Mean absolute scaled error (MASE) is described by (Hyndman & Athanasopoulos, 2021). Values closer to zero reflect greater accuracy.

The adapted formula for MASE with non-time series data is described here (https://stats.stackexchange.com/a/108963/20338)[7] (archived at https://perma.cc/G469-8NAJ). Scaled errors are calculated using Equation 14.38:

$$\begin{aligned}\text{scaled error}(q_i) &= \frac{\text{observed}_i - \text{predicted}_i}{\text{scaling factor}} \\ &= \frac{\text{observed}_i - \text{predicted}_i}{\frac{1}{n}\sum\limits_{i=1}^{n}|\text{observed}_i - \overline{\text{observed}}|}\end{aligned} \tag{14.38}$$

Then, we calculate the mean of the absolute value of the scaled errors to get MASE, as in Equation 14.39:

$$\begin{aligned}\text{mean absolute scaled error (MASE)} &= \frac{1}{n}\sum_{i=1}^{n}|q_i| \\ &= \text{mean}(|\text{scaled error}|) \\ &= \text{mean}(|q_i|)\end{aligned} \tag{14.39}$$

Note: MASE is undefined when the scaling factor is zero, due to division by zero. With non-time series data, the scaling factor is the average of the absolute value of individuals' observed scores minus the average observed score ($\frac{1}{n}\sum\limits_{i=1}^{n}|\text{observed}_i - \overline{\text{observed}}|$).

14.7.1.2.5 Root Mean Squared Log Error (RMSLE)

The squared log of the accuracy ratio is described by Tofallis (2015). The accuracy ratio is in Equation 14.40:

[6] https://github.com/DevPsyLab/petersenlab
[7] https://stats.stackexchange.com/a/108963/20338

$$\text{accuracy ratio} = \frac{\text{predicted}_i}{\text{observed}_i} \tag{14.40}$$

However, the accuracy ratio is undefined with observed or predicted values of zero, so it is common to modify it by adding 1 to the predictor and denominator, as in Equation 14.41:

$$\text{accuracy ratio} = \frac{\text{predicted}_i + 1}{\text{observed}_i + 1} \tag{14.41}$$

Squaring the log values keeps the values positive, such that smaller values (values closer to zero) reflect greater accuracy. Then we take the mean of the squared log values, which keeps the values positive, and calculate the square root of the mean squared log values to put them back on the (pre-squared) log metric. This is known as the root mean squared log error (RMSLE). Division inside the log is equal to subtraction outside the log. So, the formula can be reformulated with the subtraction of two logs, as in Equation 14.42:

$$\begin{aligned}
\text{root mean squared log error (RMSLE)} &= \sqrt{\sum_{i=1}^{n} log\left(\frac{\text{predicted}_i + 1}{\text{observed}_i + 1}\right)^2} \\
&= \sqrt{\text{mean}\left[log\left(\frac{\text{predicted}_i + 1}{\text{observed}_i + 1}\right)^2\right]} \\
&= \sqrt{\text{mean}[log(\text{accuracy ratio})^2]} = \sqrt{\text{mean}\Big\{[log(\text{predicted}_i + 1) - log}
\end{aligned} \tag{14.42}$$

RMSLE can be preferable when the scores have a wide range of values and are skewed. RMSLE can help to reduce the impact of outliers. RMSLE gives more weight to smaller errors in the prediction of small observed values, while also penalizing larger errors in the prediction of larger observed values. It overweights underestimates and underweights overestimates.

There are other variations of prediction accuracy metrics that use the log of the accuracy ratio. One variation makes it similar to median symmetric percentage error (Morley et al., 2018).

Note: Root mean squared log error is undefined when one or more predicted values or actual values equals -1. When predicted or actual values are -1, this leads to $log(0)$, which is undefined. The `accuracyOverall()` function of the `petersenlab`[8] package (Petersen, 2024a) provides the option in the function to drop undefined values so you can still generate an estimate of accuracy despite undefined values.

[8] https://github.com/DevPsyLab/petersenlab

14.7.1.2.6 Coefficient of Determination (R^2)

The coefficient of determination (R^2) reflects the proportion of variance in the outcome (dependent) variable that is explained by the model predictions: $R^2 = \frac{\text{variance explained in } Y}{\text{total variance in } Y}$. Larger values indicate greater accuracy.

R^2 is commonly estimated in multiple regression, in which multiple predictors are allowed to predict one outcome.

14.7.1.2.6.1 Adjusted R^2 (R^2_{adj})

Adjusted R^2 is similar to the coefficient of determination, but it accounts for the number of predictors included in the regression model to penalize overfitting. Adjusted R^2 reflects the proportion of variance in the outcome (dependent) variable that is explained by the model predictions over and above what would be expected to be accounted for by chance, given the number of predictors in the model. Larger values indicate greater accuracy. The formula for adjusted R^2 is in Equation 9.4. Adjusted R^2 is described further in Section 9.4.

14.7.1.2.6.2 Predictive R^2

Predictive R^2 is described here: https://tomhopper.me/2014/05/16/can-we-do-better-than-r-squared/ (archived at https://perma.cc/BK8J-HFUK). Predictive R^2 penalizes overfitting, unlike traditional R^2. Larger values indicate greater accuracy.

14.7.2 Discrimination

When dealing with a categorical outcome, discrimination is the ability to separate events from non-events. When dealing with a continuous outcome, discrimination is the strength of the association between the predictor and the outcome. Threshold-dependent aspects of discrimination at a particular cutoff (e.g., sensitivity, specificity) are described in Section 14.6.

14.7.2.1 Area under the ROC curve (AUC)

The area under the ROC curve (AUC) is a general index of discrimination accuracy for a categorical outcome. It is also called the concordance (c) statis-

tic. Larger values reflect greater discrimination accuracy. AUC was estimated using the `pROC` package (Robin et al., 2023).

14.7.2.2 Effect Size (β) of Regression

The effect size of a predictor, i.e., the standardized regression coefficient is called a beta (β) coefficient, is a general index of discrimination accuracy for a continuous outcome. Larger values reflect greater accuracy. We can obtain standardized regression coefficients by standardizing the predictors and outcome using the `scale()` function in R.

14.7.3 Calibration

When dealing with a categorical outcome, calibration is the degree to which a probabilistic estimate of an event reflects the true underlying probability of the event. When dealing with a continuous outcome, calibration is the degree to which the predicted values are close in value to the outcome values. The importance of examining calibration (in addition to discrimination) is described by Lindhiem et al. (2020). Calibration can be examined in several ways, including Spiegelhalter's z (see Section 14.7.3.2), and the mean difference between predicted and observed values at different binned thresholds as depicted graphically with a calibration plot (see Figure 14.7).

14.7.3.1 Calibration Plot

Calibration plots can be helpful for identifying miscalibration. A calibration plot depicts the predicted probability of an event on the x-axis, and the actual (observed) probability of the event on the y-axis. The predictions are binned into a certain number of groups (commonly 10). The diagonal line reflects predictions that are perfectly calibrated. To the extent that predictions deviate from the diagonal line, the predictions are miscalibrated.

Well-calibrated predictions are depicted in Figure 14.6:

```
# Specify data
examplePredictionsWellCalibrated <- seq(from = 0, to = 1, by = .1)
exampleOutcomesWellCalibrated <- seq(from = 0, to = 1, by = .1)

# Plot
plot(
  examplePredictionsWellCalibrated,
  exampleOutcomesWellCalibrated,
  xlim = c(0,1),
  ylim = c(0,1),
```

```
  xlab = "Predicted Probability",
  ylab = "Observed Proportion",
  bty = "l",
  type = "n")

lines(
  c(0,1),
  c(0,1),
  lwd = 2,
  col = "#377eb8")

points(
  examplePredictionsWellCalibrated,
  exampleOutcomesWellCalibrated,
  cex = 1.5,
  col = "#e41a1c",
  lwd = 2,
  type = "p")
```

Figure 14.6 Predictions that are Well-Calibrated. That is, the predicted values are close to the observed values.

The various types of general miscalibration are depicted in Figure 14.7:

```
# Specify data
examplePredictions <- seq(from = 0, to = 1, by = .1)
exampleOutcomes <- c(0, .15, .3, .4, .45, .5, .55, .6, .7, .85, 1)

overPrediction <- c(0, .02, .05, .1, .15, .2, .3, .4, .5, .7, 1)
underPrediction <- c(0, .3, .5, .6, .7, .8, .85, .9, .95, .98, 1)
overExtremity <- c(0, .3, .38, .42, .47, .5, .53, .58, .62, .7, 1)
underExtremity <- c(0, .05, .08, .11, .2, .5, .8, .89, .92, .95, 1)

# Plot
par(
```

```
  mfrow = c(2,2),
  mar = c(5,4,1,1) + 0.1) #margins: bottom, left, top, right

plot(
  examplePredictions,
  overExtremity,
  xlim = c(0,1),
  ylim = c(0,1),
  main = "Overextremity",
  xlab = "Predicted Probability",
  ylab = "Observed Proportion",
  bty = "l",
  cex = 1.5,
  col = "#e41a1c",
  type = "o")

lines(
  c(0,1),
  c(0,1),
  lwd = 2,
  col = "#377eb8")

plot(
  examplePredictions,
  underExtremity,
  xlim = c(0,1),
  ylim = c(0,1),
  main = "Underextremity",
  xlab = "Predicted Probability",
  ylab = "Observed Proportion",
  bty = "l",
  cex = 1.5,
  col = "#e41a1c",
  type = "o")

lines(
  c(0,1),
  c(0,1),
  lwd = 2,
  col = "#377eb8")

plot(
  examplePredictions,
  overPrediction,
```

```
  xlim = c(0,1),
  ylim = c(0,1),
  main = "Overprediction",
  xlab = "Predicted Probability",
  ylab = "Observed Proportion",
  bty = "l",
  cex = 1.5,
  col = "#e41a1c",
  type = "o")

lines(
  c(0,1),
  c(0,1),
  lwd = 2,
  col = "#377eb8")

plot(
  examplePredictions,
  underPrediction,
  xlim = c(0,1),
  ylim = c(0,1),
  main = "Underprediction",
  xlab = "Predicted Probability",
  ylab = "Observed Proportion",
  bty = "l",
  cex = 1.5,
  col = "#e41a1c",
  type = "o")

lines(
  c(0,1),
  c(0,1),
  lwd = 2,
  col = "#377eb8")
```

Figure 14.7 Types of Miscalibration. From Petersen (2024b) and Petersen (2024c).

However, predictions could also be miscalibrated in more specific ways. For instance, predictions could be well-calibrated at all predicted probabilities except for a given predicted probability (e.g., 20%). Or, the predictions could be miscalibrated but not systematically over- or underpredicted. Thus, it is important to evaluate a calibration plot to evaluate the extent to which the predictions are miscalibrated and the pattern of that miscalibration.

14.7.3.2 Spiegelhalter's z

Spiegelhalter's z was calculated using the `rms` package (Harrell, Jr., 2024). Smaller z values (and larger associated p-values) reflect greater calibration accuracy. A statistically significant Spiegelhalter's z ($p < .05$) indicates a

significant degree of miscalibration.

14.7.3.3 Calibration for predicting a continuous outcome

When predicting a continuous outcome, calibration of the predicted values in relation to the outcome values can be examined in multiple ways including:

- in a calibration plot, the extent to which the intercept is near zero and the slope is near one
- in a calibration plot, the extent to which the 95% confidence interval of the observed value, across all values of the predicted values, includes the diagonal reference line with an intercept of zero and a slope of one
- mean error
- mean absolute error
- mean squared error
- root mean squared error

With a plot of the predictions on the x-axis, and the outcomes on the y-axis (i.e., a calibration plot), calibration can be examined graphically as the extent to which the best-fit regression line has an intercept (alpha) close to zero and a slope (beta) close to one (Stevens & Poppe, 2020; Steyerberg & Vergouwe, 2014). The intercept is also called "calibration-in-the-large", whereas "calibration-in-the-small" refers to the extent to which the predicted values match the observed values at a specific predicted value (e.g., when the weather forecaster says that there is a 10% chance of rain, does it actually rain 10% of the time?). For predictions to be well calibrated, the intercept should be close to zero and the slope should be close to one. If the slope is close to one but the intercept is not close to zero (or the intercept is close to zero but the slope is not close to one), the predictions would not be considered well calibrated. The 95% confidence interval of the observed value, across all values of the predicted values, should include the diagonal reference line whose intercept is zero and whose slope is one.

For instance, based on the intercept and slope of the calibration plot in Figure INSERT, the predictions are not well calibrated, despite having a slope near one, because the 95% confidence interval of the intercept does not include zero. The best-fit line is the yellow line. The intercept from the best-fit line is positive, as shown in the regression equation. This is a case of underprediction, where the predicted values are consistently less than the observed values. The confidence interval of the observed value (i.e., the purple band) is the interval within which we have 95% confidence that the true observed value would lie for a given predicted value, based on the model The 95% prediction interval of the observed value (i.e., the dashed red lines) is the interval within which we would expect that 95% of future observations would lie for a given predicted value. The black diagonal line indicates the reference line with an intercept of

zero and a slope of one. The predictions would be significantly miscalibrated at a given level of the predicted values if the 95% confidence interval of the observed value does not include the reference line at that level of the predicted value. In this case, the 95% confidence interval of the observed value does not include the reference line (i.e., the actual observed value) at lower levels of the predicted values, so the predictions are miscalibrated lower levels of the predicted values.

Gold-standard recommendations include examining the predicted values in relation to the observed values using locally estimated scatterplot smoothing (LOESS) (Austin & Steyerberg, 2014), such as in Figure INSERT. We can examine whether the LOESS-based 95% confidence interval of the observed value at every level of the predicted values includes the diagonal reference line (i.e., the actual observed value). In this case, the 95% confidence interval of the observed value does not include the reference line at lower levels of the predicted values, so the predictions are miscalibrated at lower levels of the predicted values.

14.8 Integrating the Accuracy Indices

After computing the accuracy indices of discrimination and (2) calibration, it is then the task to integrate the indices to determine (a) which are the most accurate predictions for the given goals, and (b) whether additional improvements and refinements to the predictions need to be made. Each of the accuracy indices is computed differently and thus reward (and penalize) predictive (in)accuracy differently. Sometimes, the the accuracy indices will paint a consistent picture regarding which predictions are the most accurate. Other times, the accuracy indices may disagree about which predictions are most accurate.

In fantasy football, when evaluating the accuracy of seasonal projections, we care most about accurately distinguishing between higher levels of points (e.g., 200 vs 150) as opposed to lower levels of points (e.g., 0 vs 10). Thus, it can be helpful to punish larger errors more heavily than smaller errors, as RMSE (unlike MAE).

Thus, we would emphasize the following metrics:

- discrimination:
  - adjusted R^2
- calibration:

  - calibration plot

- general accuracy:
  - RMSE

If you focus on only one accuracy index, RMSE would be a good choice. However, I would also examine a calibration plot to evaluate whether predictions are poorly calibrated at higher levels of points. I would also examine ME—not to compare the accuracy of various predictions per se—but to determine whether predictions are systematically under- or overestimating actual points. If so, predictions may be able to be refined by adding or subtracting a constant to the predictions; however, this could worsen other accuracy indices, so it is important to conduct an iterative process of modifying then evaluating, then further modifiying and evaluating, etc. It may also be valuable to evaluate the accuracy of various subsets of the predictions. For instance, you might examine the predictive accuracy of players whose project points are greater than 100, to evaluate the accuracy of predictions specifically to distinguish between players at higher levels of points.

If we are making predictions about a categorical variable, we would emphasize the following metrics:

- discrimination:
  - area under the receiver operating curve
  - and, secondarily—depending on the particulary cutoff and the relative costs of false positives versus false negatives:
    * sensitivity
    * specificicity
    * positive predictive value
    * negative predictive value
- calibration:
  - calibration plot
  - Spiegelhalter's z
  - Mean difference between observed and predicted values

14.9 Theory Versus Empiricism

One question that inevitably arises when making predictions is the extent to which one should leverage theory versus empiricism. Theory involves conceptual claims of understanding how the causal system works (i.e., what influences

what). For example, use of theory in prediction might involve specification of the causal system that influences player performance, measurement of those factors, and the integration of that information to make a prediction. Empiricism involves "letting the data speak for themselves" and is an atheoretical approach. For example, empiricism might involve examining how thousands of variables are associated with the criterion of interest (e.g., fantasy points) and developing the best-fitting model based on those thousands of predictor variables.

Although the atheoretical approach can perform reasonably well, it can be improved by making better use of theory. An empirical result (e.g., a correlation) might not necessarily have a lot of meaning associated with it. As the maxim goes, correlation does not imply causation. Moreover, empiricism can lead to overfitting. So, empiricism is often not enough.

As Silver (2012) notes, "The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning." (p. 9). If we *understand* the variables in the system and how they influence each other, we can predict things more accurately than predicting for the sake of predicting. For instance, we have made great strides in the last decades when it comes to more accurate weather forecasts[9] (archived at https://perma.cc/PF8P-BT3D), including extreme weather events like hurricanes. These great strides have more to do with a better causal understanding of the weather system and the ability to conduct simulations of the atmosphere than merely because of big data (Silver, 2012). By contrast, other events are still incredibly difficult to predict, including earthquakes, in large part because we do not have a strong understanding of the system (and because we do not have ways of precisely measuring those causes because they occur at a depth below which we are realistically able to drill) (Silver, 2012).

At the same time, in the social and behavioral sciences, our theories of the causal processes that influence outcomes are not yet very strong. Indeed, I have misgivings calling them theories because they do not meet the traditional scientific standard for a theory. A scientific theory is an explanation of the natural world that is testable and falsifiable, and that has withstood rigorous scientific testing and scrutiny. In psychology (and other areas of social and behavioral sciences), our "theories" are more like conceptual frameworks. And these conceptual frameworks are often vague, do not make specific predictions of effects *and* noneffects, and do not hold up consistently when rigorously tested. As described by Meehl (1978):

> I consider it unnecessary to persuade you that most so-called

[9] https://www.npr.org/sections/money/2023/07/11/1186458991/should-we-invest-more-in-weather-forecasting-it-may-save-your-life

> "theories" in the soft areas of psychology (clinical, counseling, social, personality, community, and school psychology) are scientifically unimpressive and technologically worthless … Perhaps the easiest way to convince yourself is by scanning the literature of soft psychology over the last 30 years and noticing what happens to theories. Most of them suffer the fate that General MacArthur ascribed to old generals—They never die, they just slowly fade away. In the developed sciences, theories tend either to become widely accepted and built into the larger edifice of well-tested human knowledge or else they suffer destruction in the face of recalcitrant facts and are abandoned, perhaps regretfully as a "nice try." But in fields like personology and social psychology, this seems not to happen. There is a period of enthusiasm about a new theory, a period of attempted application to several fact domains, a period of disillusionment as the negative data come in, a growing bafflement about inconsistent and unreplicable empirical results, multiple resort to ad hoc excuses, and then finally people just sort of lose interest in the thing and pursue other endeavors. (pp. 806–807).

Even if we had strong theoretical understanding of the causal system that influences behavior, we would likely still have difficulty making accurate predictions because the field has largely relied on relatively crude instruments. According to one philosophical perspective known as LaPlace's demon, if we were able to know the exact conditions of everything in the universe, we would be able to know how the conditions would be in the future. This is an example of scientific determinism, where if you know the initial conditions, you also know the future. Other perspectives, such as quantum mechanics and chaos theory, would say that, even if we knew the initial conditions with 100% certainty, there would still be uncertainty in our understanding of the future. But assume, for a moment, that LaPlace's demon is true. A challenge in the social and behavioral sciences is that we have a relatively poor understanding of the initial conditions of the universe. Thus, our predictions would necessarily be probabilistic, similar to weather forecasts. Despite having a strong understanding of how weather systems behave, we have imperfect understanding of the initial conditions (e.g., the position and movement of all molecules) (Silver, 2012).

Theories tend to make grand conceptual claims that one observed variable influences another observed variable through a complex chain of intervening processes that are unobservable. Empiricism provides rich lower-level information, but lacks the broader picture. So, it seems, that we need both theory *and* empiricism. Theory and empiricism can—and should—inform each other.

14.10 Test Bias

Test bias refers to systematic error (in measurement, prediction, etc.) as a function of group membership that leads the same score to have different meaning for different groups. For instance, if the Wonderlic Contemporary Cognitive Ability Test is a strong predictor of performance for Quarterbacks but not for Running Backs, the test is biased. Test bias, including how to identify and address it, is described in Petersen (2024c)[10].

14.11 Ways to Improve Prediction Accuracy

On the whole, experts' predictions are inaccurate. Experts' predictions from many different domains tend to be inaccurate, including political scientists (Tetlock, 2017), physicians (Koehler et al., 2002), clinical psychologists (Oskamp, 1965), stock market traders and corporate financial officers (Skala, 2008), seismologists' predictions of earthquakes (Hough, 2016), economists' predictions about the economy (Makridakis et al., 2009), lawyers (Koehler et al., 2002), and business managers (Russo & Schoemaker, 1992). The most common pattern of experts' predictions is that they show overextremity, that is, their predictions have probability judgments that tend to be too extreme, as described in Section Section 14.3.2. Overextremity of experts' predictions reflects the overprecision type of overconfidence bias. The degree of confidence of a person's predictions is often not a good indicator of the accuracy of their predictions [and confidence and prediction accuracy are sometimes inversely associated; Silver (2012)]. Heuristics such as the anchoring and adjustment heuristic, cognitive biases such as confirmation bias (Hoch, 1985; Koriat et al., 1980), fallacies such as the base rate fallacy (Eddy, 1982; Koehler et al., 2002) could contribute to overconfidence of predictions. Poorly calibrated predictions are especially likely when the base rate is very low (e.g., suicide) or when the base rate is very high (Koehler et al., 2002).

Nevertheless, there are some domains that have shown greater predictive accuracy, from which we may learn what practices may lead to greater accuracy. For instance, experts have shown stronger predictive accuracy in weather forecasting (Murphy & Winkler, 1984), horse race betting (Johnson & Bruce, 2001), and playing the card game of bridge (Keren, 1987), but see Koehler et al. (2002) for exceptions.

[10] https://isaactpetersen.github.io/Principles-Psychological-Assessment/bias.html

Here are some potential ways to improve the accuracy (and honesty) of predictions and judgments:

- Provide appropriate anchoring of your predictions to the base rate of the phenomenon you are predicting. To the extent that the base rate of the event you are predicting is low, more extreme evidence should be necessary to consistently and accurately predict that the event will occur. Applying actuarial formulas and Bayes' theorem can help you appropriately weigh the base rate and evidence.
- Include multiple predictors, ideally from different measures and measurement methods. Include the predictors with the strongest validity based on theory of the causal process and based on criterion-related validity.
- When possible, aggregate multiple perspectives of predictions, especially predictions made independently (from different people/methods/etc.). The "wisdom of the crowd" is often more accurate than individuals' predictions, including predictions by so-called "experts" (Silver, 2012).
- A goal of prediction is to capture as much signal as possible and as little noise (error) as possible (Silver, 2012). Parsimony (i.e., not having too many predictors) can help reduce the amount of error variance captured by the prediction model. However, to accurately model complex systems like human behavior, complex models may be necessary. However, strong theory of the causal processes and dynamics may be necessary to develop accurate complex models.
- Although incorporating theory can be helpful, provide more weight to empiricism than to theory, until our theories and measures are stronger. Ideally, we would use theory to design a model that mirrors the causal system, with accurate measures of each process in the system, so we could make accurate predictions. However, as described in Section 14.9, our psychological theories of the causal processes that influence behavior are not yet very strong. Until we have stronger theories that specify the causal process for a given outcome, and until we have accurate measures of those causal processes, actuarial approaches are likely to be most accurate, as discussed in Chapter 12. At the same time, keep in mind that measures involving human behavior, and their resulting data, are often noisy. As a result, theoretically (conceptually) informed empirical approaches may lead to more accuracy than empiricism alone.
- Use an empirically validated and cross-validated statistical algorithm to combine information from the predictors in a formalized way. Give each predictor appropriate weight in the statistical algorithm, according to its strength of association with the outcome. Use measures with strong reliability and validity for assessing these processes to be used in the algorithm. Cross-validation will help reduce the likelihood that your model is fitting to noise and will maximize the likelihood that the model predicts accurately when applied to new data (i.e., the model's predictions accurately generalize), as described in Section 12.6.

- When presenting your predictions, acknowledge what you do not know.
- Express your predictions in terms of probabilistic estimates and present the uncertainty in your predictions with confidence intervals [even though bolder, more extreme predictions tend to receive stronger television ratings; Silver (2012)].
- Qualify your predictions by identifying and noting counter-examples that would not be well fit by your prediction model, such as extreme cases, edge cases, and "broken leg" (Meehl, 1957) cases.
- Provide clear, consistent, and timely feedback on the outcomes of the predictions to the people making the predictions (Bolger & Önkal-Atay, 2004).
- Be self-critical about your predictions. Update your judgments based on their accuracy, rather than trying to confirm your beliefs (Atanasov et al., 2020).
- In addition to considering the accuracy of the prediction, consider the quality of the prediction *process*, especially when random chance is involved to a degree, such as in poker and fantasy football (Silver, 2012).
- Work to identify and mitigate potential blindspots; be aware of cognitive biases and fallacies, such as confirmation bias and the base rate fallacy.
- Evaluate for the possibility of test bias. Correct for any test bias.

14.12 Conclusion

When the base rate of a behavior is very low or very high, you can be highly accurate in predicting the behavior by predicting from the base rate. Thus, you cannot judge how accurate your prediction is until you know how accurate your predictions would be by random chance. Moreover, maximizing percent accuracy may not be the ultimate goal because different errors have different costs. Though there are many indices of accuracy, there are two general types of accuracy: discrimination and calibration. Discrimination accuracy is frequently evaluated with the area under the receiver operating characteristic curve, or with sensitivity and specificity, or with standardized regression coefficients or the coefficient of determination. Calibration accuracy is frequently evaluated graphically and with various indices. Sensitivity and specificity depend on the cutoff. It is important to evaluate both discrimination and calibration when evaluating prediction accuracy.

```
             v1
1 This is a Test
```

15

Mythbusters: Putting Fantasy Football Beliefs/Anecdotes to the Test

15.1 Getting Started

15.1.1 Load Packages

15.1.2 Load Data

```
load(file = here("data", "nfl_playerContracts.RData"))
```

```
Error in here("data", "nfl_playerContracts.RData"): could not find function "here"
```

```
load(file = here("data", "nfl_pbp.RData"))
```

```
Error in here("data", "nfl_pbp.RData"): could not find function "here"
```

15.2 Do Players Score More Fantasy Points in their Contract Year?

```
nfl_playerContracts_subset %>%
  filter(player == "Aaron Rodgers") %>%
  select(player:years, contractYear, next_contract_start, played_in_contract_year)
```

```
Error in nfl_playerContracts_subset %>% filter(player == "Aaron Rodgers") %>% : could not find functior
```

```
nfl_playerContracts_subset %>%
  filter(player %in% c("Jared Allen", "Aaron Rodgers")) %>%
  select(player:years, contractYear, next_contract_start, played_in_contract_year)
```

```
Error in nfl_playerContracts_subset %>% filter(player %in% c("Jared Allen", : could not find function '
```

```
# Subset to remove players without a year signed
nfl_playerContracts_subset <- nfl_playerContracts %>%
  filter(!is.na(year_signed) & year_signed != 0)
```

```
Error in nfl_playerContracts %>% filter(!is.na(year_signed) & year_signed != : could not find function
```

```
# Determine the contract year for a given contract
nfl_playerContracts_subset$contractYear <- nfl_playerContracts_subset$year_signed + nfl_playerCont
```

```
Error in eval(expr, envir, enclos): object 'nfl_playerContracts_subset' not found
```

```
# Arrange contracts by player and year_signed
nfl_playerContracts_subset <- nfl_playerContracts_subset %>%
  group_by(player, position) %>%
  arrange(player, position, -year_signed) %>%
  ungroup()
```

```
Error in nfl_playerContracts_subset %>% group_by(player, position) %>% : could not find function "%>%"
```

```
# Determine if the player played in the original contract year
nfl_playerContracts_subset <- nfl_playerContracts_subset %>%
  group_by(player, position) %>%
  mutate(
    next_contract_start = lag(year_signed)) %>%
  ungroup() %>%
  mutate(
    played_in_contract_year = ifelse(
      is.na(next_contract_start) | contractYear < next_contract_start,
      TRUE,
      FALSE))
```

```
Error in nfl_playerContracts_subset %>% group_by(player, position) %>% : could not find function "%>%"
```

```
# Merge with seasonal fantasy points data
```

15.3 Conclusion

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Andersen, D., Petersen, I. T., & Tungate, A. (2024). *ffanalytics: Scrape data for fantasy football*. https://github.com/FantasyFootballAnalytics/ffanalytics

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, *160*, 19–35. https://doi.org/10.1016/j.obhdp.2020.02.001

Austin, P. C., & Steyerberg, E. W. (2014). Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*, *33*(3), 517–535. https://doi.org/10.1002/sim.5941

Avugos, S., Köppen, J., Czienskowski, U., Raab, M., & Bar-Eli, M. (2013). The "hot hand" reconsidered: A meta-analytic approach. *Psychology of Sport and Exercise*, *14*(1), 21–27. https://doi.org/10.1016/j.psychsport.2012.07.005

Baird, C., & Wagner, D. (2000). The relative validity of actuarial- and consensus-based risk assessment systems. *Children and Youth Services Review*, *22*(11), 839–871. https://doi.org/10.1016/S0190-7409(00)00122-5

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise*, *7*(6), 525–553. https://doi.org/10.1016/j.psychsport.2006.03.001

Bocskocsky, A., Ezekowitz, J., & Stein, C. (2014). The hot hand: A new approach to an old "fallacy." *MIT Sloan Sports Analytics Conference*.

Bolger, F., & Önkal-Atay, D. (2004). The effects of feedback on judgmental interval predictions. *International Journal of Forecasting*, *20*(1), 29–39. https://doi.org/10.1016/S0169-2070(03)00009-8

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, *116*(536), 2009–2022. https://doi.org/10.1080/01621459.2020.1758115

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd

ed.). Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9780203771587

Congelio, B. J. (2023). *Introduction to NFL analytics with R*. CRC Press. https://bradcongelio.com/nfl-analytics-with-r-book

Corston, R., & Colman, A. M. (2000). *A crash course in SPSS for windows*. Wiley-Blackwell.

D'Onofrio, B. M., Sjölander, A., Lahey, B. B., Lichtenstein, P., & Öberg, A. S. (2020). Accounting for confounding in observational studies. *Annual Review of Clinical Psychology*, *16*(1), 25–48. https://doi.org/10.1146/annurev-clinpsy-032816-045030

Dana, J., & Thomas, R. (2006). In defense of clinical judgment … and mechanical prediction. *Journal of Behavioral Decision Making*, *19*(5), 413–428. https://doi.org/10.1002/bdm.537

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Den Hartigh, R. J. R., Niessen, A. S. M., Frencken, W. G. P., & Meijer, R. R. (2018). Selection procedures in sports: Improving predictions of athletes' future performance. *European Journal of Sport Science*, *18*(9), 1191–1198. https://doi.org/10.1080/17461391.2018.1480662

Digitale, J. C., Martin, J. N., & Glymour, M. M. (2022). Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*, *142*, 264–267. https://doi.org/10.1016/j.jclinepi.2021.08.001

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge University Press.

Farrington, D. P., & Loeber, R. (1989). Relative improvement over chance (RIOC) and phi as measures of predictive efficiency and strength of association in 2×2 tables. *Journal of Quantitative Criminology*, *5*(3), 201–213. https://doi.org/10.1007/BF01062737

Garb, H. N., & Wood, J. M. (2019). Methodological advances in statistical prediction. *Psychological Assessment*, *31*(12), 1456–1466. https://doi.org/10.1037/pas0000673

Getty, D., Li, H., Yano, M., Gao, C., & Hosoi, A. E. (2018). Luck and the law: Quantifying chance in fantasy sports and other contests. *SIAM Review*, *60*(4), 869–887. https://doi.org/10.1137/16m1102094

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314. https://doi.org/10.1016/0010-0285(85)90010-6

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction

procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Harrell, Jr., F. E. (2024). *rms: Regression modeling strategies*. https://hbiostat.org/R/rms/

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 719–731. https://doi.org/10.1037/0278-7393.11.1-4.719

Hough, S. E. (2016). *Predicting the unpredictable: The tumultuous science of earthquake prediction*. Princeton University Press.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3

Johnson, J. E. V., & Bruce, A. C. (2001). Calibration of subjective probability judgments in a naturalistic setting. *Organizational Behavior and Human Decision Processes*, *85*(2), 265–290. https://doi.org/10.1006/obhd.2000.2949

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, *39*(1), 98–114. https://doi.org/10.1016/0749-5978(87)90047-1

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2020). Suicide prediction models: A critical review of recent research with recommendations for the way forward. *Molecular Psychiatry*, *25*(1), 168–179. https://doi.org/10.1038/s41380-019-0531-0

Kievit, R., Frankenhuis, W., Waldorp, L., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology*, *4*(513). https://doi.org/10.3389/fpsyg.2013.00513

Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Kotrba, V. (2020). Heuristics in fantasy sports: Is it profitable to strategize based on favourite of the match? *Mind & Society*, *19*(1), 195–206. https://doi.org/10.1007/s11299-020-00231-7

Lederer, D. J., Bell, S. C., Branson, R. D., Chalmers, J. D., Marshall, R., Maslove, D. M., Ost, D. E., Punjabi, N. M., Schatz, M., Smyth, A. R., Stewart, P. W., Suissa, S., Adjei, A. A., Akdis, C. A., Azoulay, É., Bakker, J., Ballas, Z. K., Bardin, P. G., Barreiro, E., … Vincent, J.-L. (2019). Control of confounding and reporting of results in causal inference studies. Guidance for authors from editors of respiratory, sleep, and critical care journals. *Annals of the American Thoracic Society*, *16*(1), 22–28. https:

//doi.org/10.1513/AnnalsATS.201808-564PS

Lee, M. D., & Liu, S. (2022). Drafting strategies in fantasy football: A study of competitive sequential human decision making. *Judgment and Decision Making*, *17*(4), 691–719. https://doi.org/10.1017/S1930297500008901

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*(1), 53–70. https://doi.org/10.1111/j.1745-6916.2007.00029.x

Lindhiem, O., Petersen, I. T., Mentch, L. K., & Youngstrom, E. A. (2020). The importance of calibration in clinical psychology. *Assessment*, *27*(4), 840–854. https://doi.org/10.1177/1073191117752055

Lyons, B. D., Hoffman, B. J., Michel, J. W., & Williams, K. J. (2011). On the predictive efficiency of past performance and physical ability: The case of the national football league. *Human Performance*, *24*(2), 158–172. https://doi.org/10.1080/08959285.2011.555218

Makridakis, S., Hogarth, R. M., & Gaba, A. (2009). Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting*, *25*(4), 794–812. https://doi.org/10.1016/j.ijforecast.2009.05.012

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods*, *11*(4), 386–401. https://doi.org/10.1037/1082-989X.11.4.386

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, *4*(4), 268–273. https://doi.org/10.1037/h0047554

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006x.46.4.806

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, *50*(3), 370–375. https://doi.org/10.1207/s15327752jpa5003_6

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*(3), 194–216. https://doi.org/10.1037/h0048070

Miller, J. B., & Sanjurjo, A. (2014). A cold shower for the hot hand fallacy. *Innocenzo Gasparini Institute for Economic Research*. https://repec.unibocconi.it/igier/igi/wp/2014/518.pdf

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather*, *16*(1), 69–88. https://doi.org/10.1002/2017SW001669

Motz, B. (2013). Fantasy football: A touchdown for undergraduate statistics education. *Proceedings of the Games, Learning, and Society Conference*, *9.0*, 222–228. https://doi.org/10.1184/R1/6686804.v1

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meterology. *Journal of the American Statistical Association*, *79*(387), 489–500. https://doi.org/10.2307/2288395

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, *29*(3), 261–265. https://doi.org/10.1037/h0022125

Pelechrinis, K., & Winston, W. (2022). The hot hand in the wild. *PLOS ONE*, *17*(1), e0261890. https://doi.org/10.1371/journal.pone.0261890

Petersen, I. T. (2024a). *petersenlab: A collection of R functions by the Petersen Lab*. https://github.com/DevPsyLab/petersenlab

Petersen, I. T. (2024c). *Principles of psychological assessment: With applied examples in R*. University of Iowa Libraries. https://doi.org/10.25820/work.007199

Petersen, I. T. (2024b). *Principles of psychological assessment: With applied examples in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003357421

Rice, M. E., Harris, G. T., & Lang, C. (2013). Validation of and revision to the VRAG and SORAG: The Violence Risk Appraisal Guide—Revised (VRAG-R). *Psychological Assessment*, *25*(3), 951–965. https://doi.org/10.1037/a0032878

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2023). *pROC: Display and analyze ROC curves*. https://xrobin.github.io/pROC/

Russo, J. E., & Schoemaker, P. J. (1992). Managing overconfidence. *Sloan Management Review*, *33*(2), 7.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail–but some don't*. Penguin.

Skala, D. (2008). Overconfidence in psychology and finance–an interdisciplinary literature review. *Bank i Kredyt*, *4*, 33–50.

Stevens, R. J., & Poppe, K. K. (2020). Validation of clinical prediction models: What does the "calibration slope" really measure? *Journal of Clinical Epidemiology*, *118*, 93–99. https://doi.org/10.1016/j.jclinepi.2019.09.016

Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *European Heart Journal*, *35*(29), 1925–1931. https://doi.org/10.1093/eurheartj/ehu207

Tetlock, P. E. (2017). *Expert political judgment: How good is it? How can we know? - New edition*. Princeton University Press.

Textor, J., Zander, B. van der, Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. (2017). Robust causal inference using directed acyclic graphs: The R package "dagitty". *International Journal of Epidemiology*, *45*(6), 1887–1894. https://doi.org/10.1093/ije/dyw341

Tofallis, C. (2015). A better measure of relative prediction accuracy for model

selection and model estimation. *Journal of the Operational Research Society*, *66*(8), 1352–1362. https://doi.org/10.1057/jors.2014.103

Treat, T. A., & Viken, R. J. (2023). Measuring test performance with signal detection theory techniques. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (2nd ed., Vol. 1, pp. 837–858). American Psychological Association.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Ursenbach, J., O'Connell, M. E., Neiser, J., Tierney, M. C., Morgan, D., Kosteniuk, J., & Spiteri, R. J. (2019). Scoring algorithms for a computer-based cognitive screening tool: An illustrative example of overfitting machine learning approaches and the impact on estimates of classification accuracy. *Psychological Assessment*, *31*(11), 1377–1382. https://doi.org/10.1037/pas0000764

Williams, A. J., Botanov, Y., Kilshaw, R. E., Wong, R. E., & Sakaluk, J. K. (2021). Potentially harmful therapies: A meta-scientific review of evidential value. *Clinical Psychology: Science and Practice*, *28*(1), 5–18. https://doi.org/10.1111/cpsp.12331

Woodland, L. M., & Woodland, B. M. (2015). The National Football League season wins total betting market: The impact of heuristics on behavior. *Southern Economic Journal*, *82*(1), 38–54. https://doi.org/10.4284/0038-4038-2013.145

Index

correlation, 203, 204

GitHub, xv

positive likelihood ratio, 296