

Notes for 431

Thomas E. Love, Ph.D.

2022-08-25

Table of contents

Working with These Notes	17
What You'll Find Here	17
The 431 Course online	18
Setting Up R	19
R Markdown	19
R Packages	19
The <code>Love-boost.R</code> script	20
Packages Used in these Notes	20
The <code>tidyverse</code>	21
Packages Not Included in the Notes at Present	22
1 Data Science and 431	23
1.1 Data Science Project Cycle	24
1.2 Data Science and the 431 Course	24
1.3 What The Course Is and Isn't	25
I Part A. Exploring Data	27
2 The Palmer Penguins	28
2.1 Setup: Packages Used Here	28
2.2 Viewing a Data Set	29
2.3 Create <code>newpenguins</code> : Eliminating Missing Data	29
2.4 Counting Things and Making Tables	30
2.5 Creating a Scatterplot	32
2.6 Six Ways To “Improve” This Graph	34
2.7 A Little Reflection	35
3 NHANES: A First Look	37
3.1 Setup: Packages Used Here	37
3.2 The NHANES data: A First Sample	37
3.3 A Quick Numerical Summary	39
3.4 Plotting Age vs. Height	39
3.5 Restriction to Complete Cases	41
3.6 The Distinction between <code>Gender</code> and <code>Sex</code>	41

3.7	Age-Height by Sex?	42
3.7.1	Can we show the Female and Male relationships in separate panels?	43
3.7.2	Can we add a smooth curve to show the relationship in each plot?	44
3.7.3	What if we want to assume straight line relationships?	45
3.8	Combining Plots with <code>patchwork</code>	46
3.9	Coming Up	48
4	Data Structures, Variable Types & Sampling NHANES	49
4.1	Setup: Packages Used Here	49
4.2	Data require structure and context	49
4.3	Sampling Adults in NHANES	50
4.3.1	Creating a Temporary, Cleaner Data Frame	50
4.3.2	Sampling <code>nh_temp</code> to obtain our <code>nh_adult750</code> sample	52
4.3.3	Summarizing the Data's Structure	53
4.3.4	What are the variables?	54
4.4	Quantitative Variables	56
4.4.1	A look at BMI (Body-Mass Index)	58
4.5	Qualitative (Categorical) Variables	58
4.6	Counting Missing Values	59
4.7	<code>nh_adults500cc</code> : A Sample of Complete Cases	62
4.8	Saving our Samples in <code>.Rds</code> files	63
5	Visualizing NHANES Data	65
5.1	Setup: Packages Used Here	65
5.2	Loading in the “Complete Cases” Sample	65
5.3	Distribution of Heights	65
5.3.1	Changing a Histogram’s Fill and Color	67
5.3.2	Using a frequency polygon	68
5.3.3	Using a dotplot	69
5.4	Height and Sex	70
5.4.1	A Boxplot of Height by Sex	72
5.4.2	Adding a violin plot	73
5.4.3	Histograms of Height by Sex	76
5.5	Looking at Pulse Rate	78
5.5.1	Pulse Rate and Physical Activity	80
5.5.2	Pulse by Sleeping Trouble	82
5.5.3	Pulse and HealthGen	83
5.5.4	Pulse Rate and Systolic Blood Pressure	84
5.5.5	Sleep Trouble vs. No Sleep Trouble?	84
5.6	General Health Status	86
5.6.1	Bar Chart for Categorical Data	87
5.6.2	Two-Way Tables	90
5.6.3	SBP by General Health Status	92

5.6.4	SBP by Physical Activity and General Health Status	93
5.6.5	SBP by Sleep Trouble and General Health Status	94
5.7	Conclusions	94
6	Summarizing Quantities	96
6.1	Setup: Packages Used Here	96
6.2	Working with the <code>nh_750</code> data	96
6.3	The <code>summary</code> function for Quantitative data	97
6.4	Measuring the Center of a Distribution	98
6.4.1	The Mean and The Median	98
6.4.2	Dealing with Missingness	100
6.4.3	The Mode of a Quantitative Variable	101
6.5	Measuring the Spread of a Distribution	102
6.5.1	The Range and the Interquartile Range (IQR)	102
6.5.2	The Variance and the Standard Deviation	104
6.5.3	Obtaining the Variance and Standard Deviation in R	104
6.5.4	Defining the Variance and Standard Deviation	105
6.5.5	Interpreting the SD when the data are Normally distributed	105
6.5.6	Chebyshev's Inequality: One Interpretation of the Standard Deviation .	107
6.6	Measuring the Shape of a Distribution	107
6.6.1	Multimodal vs. Unimodal distributions	107
6.6.2	Skew	108
6.6.3	Kurtosis	109
6.7	Numerical Summaries for Quantitative Variables	111
6.7.1	<code>favstats</code> in the <code>mosaic</code> package	111
6.7.2	<code>describe</code> in the <code>psych</code> package	113
6.7.3	The <code>Hmisc</code> package's version of <code>describe</code>	114
6.7.4	Other options	115
7	Summarizing Categories	116
7.1	Setup: Packages Used Here	116
7.2	Using the <code>nh_adult750</code> data again	116
7.3	The <code>summary</code> function for Categorical data	117
7.4	Tables to describe One Categorical Variable	117
7.5	Constructing Tables Well	119
7.5.1	Alabama First!	119
7.5.2	ALL is different and important	120
7.6	The Mode of a Categorical Variable	120
7.7	<code>describe</code> in the <code>Hmisc</code> package	121
7.8	Cross-Tabulations of Two Variables	123
7.9	Cross-Classifying Three Categorical Variables	127
7.10	Gaining Control over Tables in R: the <code>gt</code> package	129

8 Missing Data and Single Imputation	130
8.1 Setup: Packages Used Here	130
8.2 A Simulated Example with 15 subjects	130
8.3 Identifying missingness with <code>naniar</code> functions	131
8.4 Missing-data mechanisms	135
8.5 Options for Dealing with Missingness	135
8.6 Complete Case (and Available Case) analyses	136
8.7 Single Imputation	136
8.8 Multiple Imputation	136
8.9 Building a Complete Case Analysis	137
8.10 Single Imputation with the Mean or Mode	137
8.11 Doing Single Imputation with <code>simputation</code>	138
9 National Youth Fitness Survey	141
9.1 Setup: Packages Used Here	141
9.2 What is the NHANES NYFS?	141
9.3 The Variables included in <code>nnyfs</code>	142
9.3.1 From the NNYFS Demographic Component	142
9.3.2 From the NNYFS Dietary Component	142
9.3.3 From the NNYFS Examination Component	143
9.3.4 From the NNYFS Questionnaire Component	143
9.4 Looking over A Few Variables	145
9.4.1 <code>SEQN</code>	147
9.4.2 <code>sex</code>	147
9.4.3 <code>age_child</code>	148
9.4.4 <code>race_eth</code>	149
9.4.5 <code>income_pov</code>	150
9.4.6 <code>bmi</code>	152
9.4.7 <code>bmi_cat</code>	153
9.4.8 <code>waist</code>	154
9.4.9 <code>triceps_skinfold</code>	155
9.5 Additional Numeric Summaries	156
9.5.1 The Five Number Summary, Quantiles and IQR	156
9.6 Additional Summaries from <code>favstats</code>	158
9.7 The Histogram	158
9.7.1 Freedman-Diaconis Rule to select bin width	159
9.7.2 A Note on Colors	161
9.8 The Frequency Polygon	163
9.9 Plotting the Probability Density Function	164
9.10 The Boxplot	165
9.10.1 Drawing a Boxplot for One Variable in <code>ggplot2</code>	165
9.10.2 About the Boxplot	166
9.11 A Simple Comparison Boxplot	167

9.12 Using <code>describe</code> in the <code>psych</code> library	170
9.12.1 The Trimmed Mean	171
9.12.2 The Median Absolute Deviation	171
9.13 Assessing Skew	171
9.13.1 Non-parametric Skewness	172
9.14 Assessing Kurtosis (Heavy-Tailedness)	172
9.14.1 The Standard Error of the Sample Mean	173
9.15 The <code>describe</code> function in the <code>Hmisc</code> package	173
9.16 Summarizing data within subgroups	175
9.17 Another Example	177
9.18 Boxplots to Relate an Outcome to a Categorical Predictor	179
9.18.1 Augmenting the Boxplot with the Sample Mean	181
9.19 Building a Violin Plot	182
9.19.1 Adding Notches to a Boxplot	184
9.20 Using Multiple Histograms to Make Comparisons	187
9.21 Using Multiple Density Plots to Make Comparisons	188
9.22 A Ridgeline Plot	191
9.23 What Summaries to Report	194
10 Assessing Normality	195
10.1 Setup: Packages Used Here	195
10.2 Introduction	195
10.3 Empirical Rule Interpretation of the Standard Deviation	196
10.4 Describing Outlying Values with Z Scores	197
10.4.1 Fences and Z Scores	197
10.5 Comparing a Histogram to a Normal Distribution	197
10.5.1 Histogram of <code>energy</code> with Normal model (with Counts)	198
10.6 Does a Normal model work well for the <code>waist</code> circumference?	200
10.7 The Normal Q-Q Plot	202
10.8 Interpreting the Normal Q-Q Plot	202
10.8.1 Data from a Normal distribution shows up as a straight line in a Normal Q-Q plot	203
10.8.2 Skew is indicated by monotonic curves in the Normal Q-Q plot	204
10.8.3 Direction of Skew	207
10.8.4 Outlier-proneness is indicated by “s-shaped” curves in a Normal Q-Q plot	207
10.9 Can a Normal Distribution Fit the <code>nnyfs</code> <code>energy</code> data Well?	211
10.10 The Ladder of Power Transformations	215
10.11 Using the Ladder	216
10.12 Protein Consumption in the NNYFS data	216
10.12.1 Using <code>patchwork</code> to compose plots	218
10.13 Can we transform the <code>protein</code> data?	219
10.13.1 The Square Root	219
10.13.2 The Logarithm	221

10.13.3 This course uses Natural Logarithms, unless otherwise specified	223
10.14 What if we considered all 9 available transformations?	223
10.15 A Simulated Data Set	226
10.16 What if we considered all 9 available transformations?	230
11 Straight Line Models	233
11.1 Setup: Packages Used Here	233
11.2 Assessing A Scatterplot	233
11.2.1 Highlighting an unusual point	234
11.2.2 Adding a Scatterplot Smooth using loess	236
11.2.3 What Line Does R Fit?	238
11.3 Correlation Coefficients	240
11.4 The Pearson Correlation Coefficient	241
11.5 Studying Correlation through 6 Examples	241
11.5.1 Data Set Alex	242
11.5.2 Data Set Bonnie	245
11.5.3 Correlations for All Six Data Sets in the Correx1 Example	247
11.5.4 Data Set Colin	248
11.5.5 Draw the Picture!	248
11.6 Estimating Correlation from Scatterplots	250
11.7 The Spearman Rank Correlation	254
11.7.1 Spearman Formula	255
11.7.2 Comparing Pearson and Spearman Correlations	255
11.7.3 Spearman vs. Pearson Example 1	255
11.7.4 Spearman vs. Pearson Example 2	257
11.7.5 Spearman vs. Pearson Example 3	258
11.7.6 Spearman vs. Pearson Example 4	259
12 Linearizing Transformations	261
12.1 Setup: Packages Used Here	261
12.2 “Linearize” The Association between Quantitative Variables	261
12.3 The Box-Cox Plot	261
12.3.1 A Few Caveats	262
12.4 A Simulated Example	262
12.5 Checking on a Transformation or Re-Expression	265
12.5.1 Checking the Correlation Coefficients	266
12.5.2 Using the <code>testTransform</code> function	266
12.5.3 Comparing the Residual Plots	267
12.6 An Example from the NNYFS data	268
12.6.1 Pearson correlation and scatterplot	269
12.6.2 Plotting the Residuals	270
12.6.3 Using the Box-Cox approach to identify a transformation	272
12.6.4 Plots after Inverse Transformation	272

13 Studying Crab Claws	275
13.1 Setup: Packages Used Here	275
13.2 The Data	275
13.3 Association of Size and Force	278
13.4 The <code>loess</code> smooth	280
13.4.1 Smoothing within Species	282
13.5 Fitting a Linear Regression Model	284
13.6 Is a Linear Model Appropriate?	286
13.6.1 The log-log model	287
13.6.2 How does this compare to our original linear model?	288
13.7 Making Predictions with a Model	289
13.7.1 Predictions After a Transformation	290
13.7.2 Comparing Model Predictions	291
14 Dehydration Recovery	293
14.1 Setup: Packages Used Here	293
14.2 The Data	293
14.3 A Scatterplot Matrix	294
14.4 Are the recovery scores well described by a Normal model?	295
14.5 Simple Regression: Using Dose to predict Recovery	297
14.6 The Scatterplot, with fitted Linear Model	297
14.7 The Fitted Linear Model	298
14.7.1 Confidence Intervals	298
14.8 Coefficient Plots	299
14.9 The Summary Output	300
14.9.1 Model Specification	301
14.9.2 Residual Summary	301
14.9.3 Coefficients Output	302
14.9.4 Correlation and Slope	302
14.9.5 Coefficient Testing	303
14.9.6 Summarizing the Quality of Fit	304
14.9.7 ANOVA F test	306
14.10 Viewing the complete ANOVA table	306
14.11 Using <code>glance</code> to summarize the model's fit	307
14.12 Plotting Residuals vs. Fitted Values	308
15 The WCGS	311
15.1 Setup: Packages Used Here	311
15.2 The Western Collaborative Group Study (<code>wcgs</code>) data set	311
15.2.1 Structure of <code>wcgs</code>	312
15.2.2 Codebook for <code>wcgs</code>	313
15.2.3 Quick Summary	314
15.3 Are the SBPs Normally Distributed?	315

15.4 Identifying and Describing SBP outliers	318
15.5 Does Weight Category Relate to SBP?	320
15.6 Re-Leveling a Factor	320
15.6.1 SBP by Weight Category	321
15.7 Are Weight and SBP Linked?	323
15.8 SBP and Weight by Arcus Senilis groups?	324
15.9 Linear Model for SBP-Weight Relationship: subjects without Arcus Senilis . .	326
15.10Linear Model for SBP-Weight Relationship: subjects with Arcus Senilis . . .	327
15.11Including Arcus Status in the model	329
15.12Predictions from these Linear Models	330
15.13Scatterplots with Facets Across a Categorical Variable	331
15.14Scatterplot and Correlation Matrices	331
15.14.1 Displaying a Correlation Matrix	333
15.14.2 Using the GGally package	333
II Part B. Comparing Summaries	335
16 Confidence Intervals for a Mean	336
16.1 Setup: Packages Used Here	336
16.2 Introduction	336
16.3 This Chapter's Goals	337
16.4 Serum Zinc Levels in 462 Teenage Males (serzinc)	337
16.5 Our Goal: A Confidence Interval for the Population Mean	338
16.6 Exploratory Data Analysis for Serum Zinc	338
16.6.1 Graphical Summaries	338
16.6.2 Numerical Summaries	340
16.7 Defining a Confidence Interval	341
16.8 Estimating the Population Mean from the Serum Zinc data	341
16.9 Confidence vs. Significance Level	342
16.10The Standard Error of a Sample Mean	343
16.11The t distribution and CIs for a Mean	343
16.11.1 The Formula	344
16.11.2 Student's t distribution	344
16.12Building the CI in R	345
16.13Using an intercept-only regression model	345
16.14Interpreting the Result	347
16.15What if we want a 95% or 99% confidence interval instead?	348
16.16Using the broom package with the t test	348
16.16.1 Effect of Changing the Confidence Level	349
16.17One-sided vs. Two-sided Confidence Intervals	349
16.18Bootstrap Confidence Intervals	351
16.19Resampling is A Big Idea	351

16.20	When is a Bootstrap Confidence Interval Reasonable?	352
16.21	Bootstrap confidence interval for the mean: Process	352
16.22	Using R to estimate a bootstrap CI	352
16.23	Comparing Bootstrap and T-Based Confidence Intervals	353
16.23.1	Bootstrap Resampling: Advantages and Caveats	354
16.24	Using the Bootstrap to develop other CIs	354
16.24.1	Changing the Confidence Level	354
16.25	One-Tailed Bootstrap Confidence Intervals	355
16.25.1	Bootstrap CI for the Population Median	355
16.25.2	Bootstrap CI for the IQR	356
16.26	Wilcoxon Signed Rank Procedure for CIs	357
16.26.1	What is a Pseudo-Median?	357
16.27	Wilcoxon Signed Rank-based CI in R	358
16.27.1	Interpreting the Wilcoxon CI for the Population Median	358
16.27.2	Using the <code>broom</code> package with the Wilcoxon test	358
16.28	General Advice	359
17	Ibuprofen in Sepsis	360
17.1	Setup: Packages Used Here	360
17.2	The Trial	360
17.3	Comparing Two Groups	362
17.3.1	Model-Based Comparisons and ANOVA/Regression	362
17.4	Key Questions for Comparing with Independent Samples	363
17.4.1	What is the population under study?	363
17.4.2	What is the sample ? Is it representative of the population?	363
17.4.3	Who are the subjects / individuals within the sample?	363
17.4.4	What data are available on each individual?	363
17.4.5	RCT Caveats	364
17.5	Exploratory Data Analysis	364
17.6	Estimating the Difference in Population Means	367
17.7	t-based CI for population mean1 - mean2 difference	368
17.7.1	The Pooled t procedure	368
17.7.2	Using linear regression to obtain a pooled t confidence interval	369
17.7.3	The Welch t procedure	371
17.8	Wilcoxon-Mann-Whitney “Rank Sum” CI	372
17.9	Bootstrapping: A More Robust Approach	373
17.9.1	Bootstrap CI for the Sepsis study	374
17.10	Summary: Specifying A Two-Sample Study Design	374
17.11	Results for the <code>sepsis</code> study	375
17.11.1	Sepsis Estimation Results	377
17.12	Categorizing the Outcome and Comparing Rates	379
17.13	Estimating the Difference in Proportions	380

18 Comparing Means with Paired Samples	381
18.1 Setup: Packages Used Here	381
18.2 Lead in the Blood of Children	381
18.3 The Lead in the Blood of Children Study	382
18.3.1 Our Key Questions for a Paired Samples Comparison	383
18.3.2 Lead Study Caveats	384
18.4 Exploratory Data Analysis for Paired Samples	384
18.4.1 The Paired Differences	385
18.4.2 Impact of Matching - Scatterplot and Correlation	387
18.5 Looking at Separate Samples: Using <code>pivot_longer</code>	388
18.6 Estimating the Difference in Means with Paired Samples	391
18.6.1 Paired Data in Longer Format?	392
18.7 Matched Pairs vs. Two Independent Samples	393
18.8 Estimating the Population Mean of the Paired Differences	394
18.9 t-based CI for Population Mean of Paired Differences	394
18.9.1 Method 1	394
18.9.2 Method 2	395
18.9.3 Method 3	396
18.9.4 Method 4	397
18.9.5 Method 5	397
18.9.6 Assumptions	398
18.10 Bootstrap CI for mean difference using paired samples	398
18.10.1 Assumptions	399
18.11 Wilcoxon Signed Rank-based CI for paired samples	399
18.11.1 Assumptions	400
18.12 Choosing a Confidence Interval Approach	401
18.13 Conclusions for the <code>bloodlead</code> study	401
18.14 The Sign test	401
18.15 Paired (Dependent) vs. Independent Samples	403
18.15.1 Three “Tricky” Examples	404
18.16 A More Complete Decision Support Tool: Comparing Means	405
18.16.1 Answers for the Three “Tricky” Examples	405
19 Hypothesis Testing: What is it good for?	407
19.1 Setup: Package Used Here	407
19.2 Introduction	407
19.3 Five Steps in any Hypothesis Test	408
19.4 Type I and Type II Error	408
19.5 The Courtroom Analogy	409
19.6 Significance vs. Importance	409
19.7 What does Dr. Love dislike about p values and “statistical significance”?	409
19.8 The ASA Articles in 2016 and 2019 on Statistical Significance and P-Values	410
19.9 Errors in Hypothesis Testing	411

19.10	The Two Types of Hypothesis Testing Errors	412
19.11	The Significance Level is the Probability of a Type I Error	412
19.12	The Probability of avoiding a Type II Error is called Power	412
19.13	Incorporating the Costs of Various Types of Errors	413
19.14	Power and Sample Size Considerations	414
19.15	Sample Size in a One-Sample t test	414
19.15.1	A Toy Example	415
19.15.2	Using the <code>power.t.t.test</code> function	415
19.16	Changing Assumptions	416
19.16.1	Increasing Sample Size Increases Power	416
19.16.2	Increasing Effect Size will increase Power	417
19.16.3	Decreasing the Standard Deviation will increase Power	418
19.16.4	Larger Significance Level increases Power	419
19.17	Paired Sample t Tests and Power/Sample Size	420
19.17.1	A Toy Example	420
19.17.2	Using the <code>power.t.t.test</code> function	420
19.17.3	Changing Assumptions in a Power Calculation	421
19.17.4	Changing the Sample Size	421
19.17.5	Changing the Effect Size	422
19.17.6	Changing the Standard Deviation	423
19.17.7	Changing the Significance Level	424
19.18	Two Independent Samples: Power for t Tests	425
19.19	A New Example	425
19.19.1	Another Scenario	426
19.20	Power for Independent Sample T tests with Unbalanced Designs	427
19.20.1	The most efficient design for an independent samples comparison will be balanced.	427
20	Two Examples Comparing Means	429
20.1	Setup: Packages Used Here	429
20.2	A Study of Battery Life	429
20.2.1	Question 1. What is the outcome under study?	430
20.2.2	Question 2. What are the treatment/exposure groups?	430
20.2.3	Question 3. Are the data collected using paired or independent samples?	430
20.2.4	Question 4. Are the data a random sample from the population of interest?	431
20.2.5	Question 5. What significance level will we use?	431
20.2.6	Question 6. Are we using a one-sided or two-sided comparison?	431
20.2.7	Question 9. What does the distribution of outcomes in each group tell us?	431
20.2.8	Inferential Results for the Battery Study	433
20.2.9	Paired Samples Approaches	434
20.2.10	Independent Samples Approaches	434
20.3	The Breakfast Study: Does Oat Bran Cereal Lower Serum LDL Cholesterol?	434
20.3.1	Question 1. What is the outcome under study?	435

20.3.2 Question 2. What are the treatment/exposure groups?	435
20.3.3 Question 3. Are the data collected using paired or independent samples?	435
20.3.4 Question 4. Are the data a random sample from the population of interest?	435
20.3.5 Question 5. What significance level will we use?	436
20.3.6 Question 6. Are we using a one-sided or two-sided comparison?	436
20.3.7 Question 7. Did pairing help reduce nuisance variation?	436
20.3.8 Question 8. What does the distribution of paired differences tell us?	436
20.4 Power, Sample Size and the Breakfast Study	438
20.4.1 The Setup	438
20.4.2 The R Calculations	438
20.4.3 Independent samples, instead of paired samples?	439
21 Estimating a Population Proportion	441
21.1 Setup: Packages Used Here	441
21.2 A First Example: Serum Zinc in the “Normal” Range?	441
21.2.1 Using an Intercept-Only Regression Again?	442
21.2.2 A $100(1-\alpha)\%$ Confidence Interval for a Population Proportion	443
21.3 Using <code>binom.test</code> from the <code>mosaic</code> package	444
21.3.1 The Wald test approach	444
21.3.2 The Clopper-Pearson approach	446
21.3.3 The Score approach	447
21.3.4 The Agresti-Coull Approach	448
21.3.5 The “Plus 4” approach	449
21.3.6 SAIFS: single augmentation with an imaginary failure or success	450
21.3.7 A Function in R to Calculate the SAIFS Confidence Interval	451
21.3.8 The <code>saifs.ci</code> function in R	452
21.4 A Second Example: Ebola Mortality Rates through 9 Months of the Epidemic	453
21.4.1 Working through the Ebola Virus Disease Example	453
21.4.2 Using R to estimate the CI for our Ebola example	454
21.4.3 Plotting the Confidence Intervals for the Ebola Virus Disease Example .	455
21.4.4 What about the <code>saifs.ci()</code> result?	455
21.5 Can the Choice of Confidence Interval Method Matter?	456
22 Comparing Proportions with Two Independent Samples	458
22.1 Setup: Packages Used Here	458
22.2 A First Example: Ibuprofen and Sepsis Trial	458
22.3 Relating a Treatment to an Outcome	460
22.4 Definitions of Probability and Odds	460
22.5 Defining the Relative Risk	460
22.6 Defining the Risk Difference	461
22.7 Defining the Odds Ratio, or the Cross-Product Ratio	461
22.8 Comparing Rates in a 2x2 Table	462

22.9 The <code>twobytwo</code> function in R	462
22.9.1 Standard Epidemiological Format	463
22.9.2 Outcome Probabilities and Confidence Intervals Within the Treatment Groups	463
22.9.3 Relative Risk, Odds Ratio and Risk Difference, with Confidence Intervals	464
22.10 Estimating a Rate More Accurately: Use $(x + 2)/(n + 4)$ rather than x/n	464
22.11 A Second Example: Ebola Virus Disease Study, again	466
23 Power and Proportions	468
23.1 Setup: Packages Used Here	468
23.2 Tuberculosis Prevalence Among IV Drug Users	468
23.3 Designing a New TB Study	469
23.4 Using <code>power.prop.test</code> for Balanced Designs	469
23.5 How <code>power.prop.test</code> works	470
23.6 A Revised Scenario	470
23.7 Using the <code>pwr</code> library for Unbalanced Designs	471
23.7.1 Calculating the Effect Size h	471
23.8 Using <code>pwr.2p2n.test</code> in R	471
23.8.1 Comparison to Balanced Design	472
24 Larger Contingency Tables	474
24.1 Setup: Packages Used Here	474
24.2 A 2x3 Table: Comparing Response to Active vs. Placebo	474
24.2.1 Getting the Table into R	475
24.2.2 Manipulating the Table's presentation	475
24.3 Accuracy of Death Certificates (A 6x3 Table)	476
24.4 The Pearson Chi-Square Test of Independence	477
24.5 Three-Way Tables: A 2x2xK Table and a Mantel-Haenszel Analysis	478
24.5.1 Smoking and Mortality in the UK	479
24.5.2 The <code>whickham</code> data with age, too	480
24.5.3 Checking Assumptions: The Woolf test	483
24.5.4 The Cochran-Mantel-Haenszel Test	484
24.5.5 Without the Continuity Correction	485
25 Analysis of Variance	486
25.1 Setup: Packages Used Here	486
25.2 National Youth Fitness Survey	486
25.3 Comparing Gross Motor Quotient Scores by Income Level (3 Categories)	487
25.4 Alternative Procedures for Comparing More Than Two Means	490
25.4.1 Extending the Welch Test to > 2 Independent Samples	491
25.4.2 Extending the Rank Sum Test to > 2 Independent Samples	492
25.4.3 Can we use the bootstrap to compare more than two means?	492

25.5 The Analysis of Variance	492
25.5.1 The <code>oneway.test</code> approach	493
25.5.2 Using the <code>aov</code> approach and the <code>summary</code> function	493
25.5.3 Using the <code>anova</code> function after fitting a linear model	493
25.6 Interpreting the ANOVA Table	494
25.6.1 What are we Testing?	494
25.6.2 Elements of the ANOVA Table	494
25.6.3 The Degrees of Freedom	494
25.6.4 The Sums of Squares	495
25.6.5 The Mean Square	496
25.6.6 The F Test Statistic and p Value	496
25.7 The Residual Standard Error	497
25.8 The Proportion of Variance Explained by the Factor	497
25.9 The Regression Approach to Compare Population Means based on Independent Samples	497
25.9.1 Interpreting the Regression Output	498
25.9.2 The Full ANOVA Table	499
25.9.3 ANOVA Assumptions	499
25.10 Equivalent approach to get ANOVA Results	500
25.11 The Problem of Multiple Comparisons	500
25.11.1 The Bonferroni solution	500
25.11.2 Pairwise Comparisons using Tukey's HSD Method	501
25.11.3 Plotting the Tukey HSD results	501
25.12 What if we consider another outcome, BMI?	503
III Part C. Building Models	508
26 Multiple Regression: Introduction	509
26.1 Reminders of a few Key Concepts	509
26.2 What is important in 431?	510
Appendices	510
A Getting Data Into R	511
Using data from an R package	511
Using <code>read_rds</code> to read in an R data set	511
Using <code>read_csv</code> to read in a comma-separated version of a data file	512
Converting Character Variables into Factors	515
Converting Data Frames to Tibbles	515
For more advice	516

Working with These Notes

1. This document is broken down into multiple chapters. Use the table of contents on the left side of the screen to navigate between chapters, or use the right side to navigate within the current chapter.
2. You can also search the document, using an automated index.
3. Any of the code provided in the document can be copied to the clipboard using the Copy icon at the top right of the code block.
4. The document is not yet complete, although a more complete draft should be in place by the start of class. Newer versions will appear unpredictably throughout the semester.

What You'll Find Here

These Notes provide a series of examples using R to work through issues that are likely to come up in PQHS/CRSP/MPHP 431. What you will mostly find are brief explanations of a key idea or summary, accompanied (most of the time) by R code and a demonstration of the results of applying that code.

While these Notes share some of the features of a textbook, they are neither comprehensive nor completely original. The main purpose is to give 431 students a set of common materials on which to draw during the course. In class, we will sometimes:

- reiterate points made in this document,
- amplify what is here,
- simplify the presentation of things done here,
- use new examples to show some of the same techniques,
- refer to issues not mentioned in this document,

but what we don't do is follow these notes very precisely. We assume instead that you will read the materials and try to learn from them, just as you will attend classes and try to learn from them. We welcome feedback of all kinds on this document or anything else.

The 431 Course online

The **online** home for Dr. Love's 431 course in Fall 2022 is

<https://thomaselove.github.io/431-2022/>.

Go there for all information related to the course.

All of the code and text in these Notes is posted online as HTML, and it is also possible to download PDF and ePub versions of the document from the down arrow next to the title (Notes for 431) at the top left of this screen. All data and R code related to these notes are also available to you through [our course web site](#).

By the end of the semester, you will also have access to the Quarto files which generate everything in the document, including all of the R results. [Quarto](#) is a souped-up version of [R Markdown](#), which you will use during the semester to complete your assignments, and which I used to develop previous versions of these notes. We will demonstrate the use of R Markdown and [RStudio](#) (the “program” we use to interface with the R language) in class.

Setting Up R

These Notes make extensive use of

- the statistical [software language R](#), and
- the development environment [RStudio](#),

both of which are free, and you'll need to install them on your machine. Instructions for doing so will be found on [the course website](#).

If you need a gentle introduction, or if you're just new to R and RStudio and need to learn about them, we encourage you to take a look at <https://moderndive.com/>, which provides an introduction to statistical and data sciences via R at Ismay and Kim (2022).

R Markdown

These notes were written using [Quarto](#), which is an amplification of R Markdown (which we'll learn in 431.) [R Markdown](#), like R and RStudio and Quarto, is free and open source.

R Markdown is described as an *authoring framework* for data science, which lets you

- save and execute R code
- generate high-quality reports that can be shared with an audience

This description comes from [RStudio's introduction to R Markdown](#) which provides an overview and quick tour of what's possible with R Markdown.

Another excellent resource to learn more about R Markdown tools is the Communicate section (especially the [R Markdown chapter](#)) of Wickham and Grolemund (2022).

R Packages

At the start of each chapter that involves R code, I'll present a series of commands I run to set up R to use several packages (libraries) of functions that expand its capabilities, make a specific change to how I want R output to be displayed (that's the `comment = NA` piece) and sets the theme for most graphs to `theme_bw()`. A chunk of code like this will occur near the top of any R Markdown work.

For example, this is the setup for one of our early chapters that loads four packages.

```
knitr::opts_chunk$set(comment = NA)

library(palmerpenguins)
library(janitor)
library(knitr)
library(tidyverse)

theme_set(theme_bw())
```

You only need to install a package once, but you need to reload it (using the `library()` function) every time you start a new session. I always load the package called `tidyverse` last, since doing so avoids some annoying problems.

The Love-boost.R script

In October, when we start Part B of the course, we'll use some special R functions I've gathered for you in a script called `Love-boost`. I'll tell R about that code using the following command...

```
source("data/Love-boost.R")
```

The `Love-boost.R` script includes four functions:

- `bootdif`
- `saifs.ci`
- `twobytwo`
- `retrodesign`

Packages Used in these Notes

A complete list of all R packages we want you to install this semester (which includes some packages not included in these Notes) is maintained at [our course web site](#).

Package	Parts	Key functions in the Package
<code>arm</code>	C	—
<code>boot</code>	B	—
<code>broom</code>	A, B, C	<code>tidy</code> , <code>augment</code> , <code>glance</code>
<code>car</code>	A, C	<code>boxCox</code> , <code>powerTransform</code> , <code>testTransform</code>

Package	Parts	Key functions in the Package
<code>Epi</code>	B	<code>twoby2</code>
<code>fivethirtyeight</code>	Appendix	source of data
<code>GGally</code>	A, C	<code>ggpairs</code>
<code>ggrepel</code>	C	—
<code>ggridges</code>	A, B	—
<code>gt</code>	A	for presenting tables
<code>Hmisc</code>	A, B, C	<code>describe</code> and others
<code>janitor</code>	A, B, C	<code>tabyl</code> and others
<code>knitr</code>	A, B, C	<code>kable</code>
<code>mice</code>	C	—
<code>mosaic</code>	A, B, C	<code>favstats</code>
<code>naniar</code>	A	<code>n_miss, miss_case_table, gg_miss_var</code>
<code>NHANES</code>	A	source of data
<code>palmerpenguins</code>	A	source of data
<code>patchwork</code>	A, B, C	for combining/annotating plots
<code>psych</code>	A, B	<code>describe</code>
<code>pwr</code>	B	—
<code>rms</code>	C	—
<code>imputation</code>	A	various imputation functions
<code>tidyverse</code>	A, B, C, Appendix	dozens of functions
<code>vcd</code>	B	—

The tidyverse

The `tidyverse` package is actually a meta-package which includes the following core packages:

- `ggplot2` for creating graphics
- `dplyr` for data manipulation
- `tidyr` for creating tidy data
- `readr` for reading in rectangular data
- `purrr` for working with functions and vectors
- `tibble` for creating tibbles - lazy, surly data frames
- `stringr` for working with data strings
- `forcats` for solving problems with factors

Loading the tidyverse with `library(tidyverse)` loads those eight packages.

Installing the tidyverse also installs several other useful packages on your machine. Read more about the `tidyverse` at <https://www.tidyverse.org/>

Packages Not Included in the Notes at Present

- `devtools`
- `equatiomatic`
- `gapminder`
- `here`
- `kableExtra`
- `magrittr`
- `markdown`
- `modelsummary`
- `nhanesA`
- `rmarkdown`
- `rmdformats`
- `rstanarm`
- `sessioninfo`
- `tableone`
- `tidymodels`
- `visdat`

1 Data Science and 431

The definition of **data science** can be a little slippery. One current view of data science, is exemplified by Steven Geringer's 2014 Venn diagram.

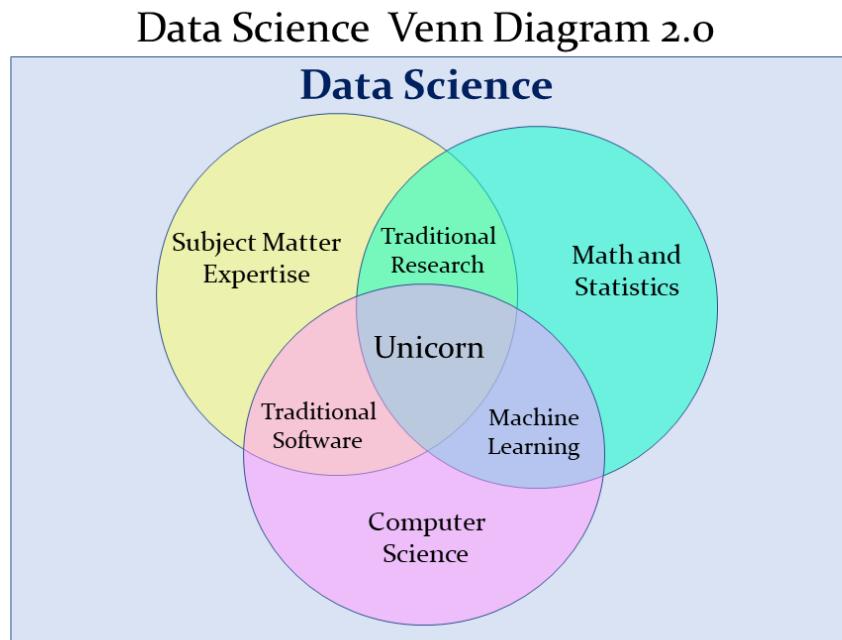


Figure 1.1: Data Science Venn Diagram from Steven Geringer

- The field encompasses ideas from mathematics and statistics and from computer science, but with a heavy reliance on subject-matter knowledge. In our case, this includes clinical, health-related, medical or biological knowledge.
- As Gelman and Nolan (2017) suggest, the experience and intuition necessary for good statistical practice are hard to obtain, and teaching data science provides an excellent opportunity to reinforce statistical thinking skills across the full cycle of a data analysis project.
- The principal form in which computer science (coding/programming) play a role in this course is to provide a form of communication. You'll need to learn how to express your ideas not just orally and in writing, but also through your code.

Data Science is a **team** activity. Everyone working in data science brings some part of the necessary skill set, but no one person can cover all three areas alone for excellent projects.

[The individual who is truly expert in all three key areas (mathematics/statistics, computer science and subject-matter knowledge) is] a mythical beast with magical powers who's rumored to exist but is never actually seen in the wild.

<http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html>

1.1 Data Science Project Cycle

A typical data science project can be modeled as follows, which comes from the introduction to the amazing book **R for Data Science**, by Garrett Grolemund and Hadley Wickham, which is a key text for this course (Wickham and Grolemund 2022).

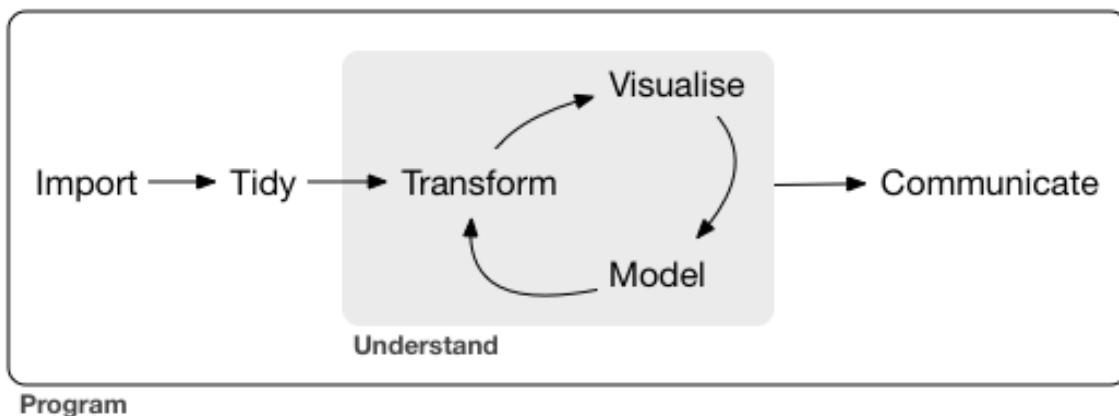


Figure 1.2: Source: R for Data Science: Introduction

This diagram is sometimes referred to as the Krebs Cycle of Data Science. For more on the steps of a data science project, we encourage you to read the Introduction of Wickham and Grolemund (2022).

1.2 Data Science and the 431 Course

We'll discuss each of these elements in the 431 course, focusing at the start on understanding our data through transformation, modeling and (especially in the early stages) visualization. In 431, we learn how to get things done.

- We get people working with R and R Studio and R Markdown, even if they are completely new to coding. A gentle introduction is provided at Ismay and Kim (2022)

- We learn how to use the **tidyverse** (<http://www.tidyverse.org/>), an array of tools in R (mostly developed by Hadley Wickham and his colleagues at R Studio) which share an underlying philosophy to make data science faster, easier, more reproducible and more fun. A critical text for understanding the tidyverse is Wickham and Grolemund (2022). Tidyverse tools facilitate:
 - **importing** data into R, which can be the source of intense pain for some things, but is really quite easy 95% of the time with the right tool.
 - **tidying** data, that is, storing it in a format that includes one row per observation and one column per variable. This is harder, and more important, than you might think.
 - **transforming** data, perhaps by identifying specific subgroups of interest, creating new variables based on existing ones, or calculating summaries.
 - **visualizing** data to generate actual knowledge and identify questions about the data - this is an area where R really shines, and we'll start with it in class.
 - **modeling** data, taking the approach that modeling is complementary to visualization, and allows us to answer questions that visualization helps us identify.
 - and last, but definitely not least, **communicating** results, models and visualizations to others, in a way that is reproducible and effective.
- Some programming/coding is an inevitable requirement to accomplish all of these aims. If you are leery of coding, you'll need to get past that, with the help of this course and our stellar teaching assistants. Getting started is always the most challenging part, but our experience is that most of the pain of developing these new skills evaporates by early October.

1.3 What The Course Is and Isn't

The 431 course is about **getting things done**. In developing this course, we adopt a modern approach that places data at the center of our work. Our goal is to teach you how to do truly reproducible research with modern tools. We want you to be able to collect and use data effectively to address questions of interest.

The curriculum includes more on several topics than you might expect from a standard graduate introduction to biostatistics.

- data gathering
- data wrangling
- exploratory data analysis and visualization
- multivariate modeling
- communication

It also nearly completely avoids formalism and is extremely applied - this is absolutely **not** a course in theoretical or mathematical statistics, and these Notes reflect that approach.

There's very little of the mathematical underpinnings here:

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

Instead, these notes (and the course) focus on how we get R to do the things we want to do, and how we interpret the results of our work. Our next Chapter provides a first example.

Part I

Part A. Exploring Data

2 The Palmer Penguins

The data in the `palmerpenguins` package in R includes information on several measurements of interest for adult foraging penguins observed on islands in the Palmer Archipelago near Palmer Station, Antarctica. Dr. Kristen Gorman and the Palmer Station Long Term Ecological Research (LTER) Program collected the data and made it available¹. The data describe three species of penguins, called Adelie, Chinstrap and Gentoo.

For more on the `palmerpenguins` package, visit <https://allisonhorst.github.io/palmerpenguins/>.

2.1 Setup: Packages Used Here

We will use the `palmerpenguins` package to supply us with data for this chapter. The `janitor` packages includes several useful functions, including `tabyl`. The `knitr` package includes the `kable()` function we'll use. Finally, the `tidyverse` package will provide the bulk of the functions we'll use in our work throughout the semester.

I always load the `tidyverse` last, because it solves some problems to do so.

```
knitr::opts_chunk$set(comment = NA)

library(palmerpenguins)
library(janitor)
library(knitr)
library(gt)
library(tidyverse)

theme_set(theme_bw())
```

¹Two fun facts: (1) Male Gentoo and Adelie penguins “propose” to females by giving them a pebble. (2) The Adelie penguin was named for his wife by Jules Dumont d’Urville, who also rediscovered the Venus de Milo.

2.2 Viewing a Data Set

The `penguins` data from the `palmerpenguins` package contains 344 rows and 8 columns. Each row contains data for a different penguin, and each column describes a variable contained in the data set.

```
penguins

# A tibble: 344 x 8
  species island   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex   year
  <fct>   <fct>           <dbl>          <dbl>            <dbl>        <dbl> <fct> <int>
1 Adelie   Torgersen      39.1          18.7             181       3750 male   2007
2 Adelie   Torgersen      39.5          17.4             186       3800 female 2007
3 Adelie   Torgersen      40.3          18               195       3250 female 2007
4 Adelie   Torgersen       NA            NA              NA        NA <NA>  2007
5 Adelie   Torgersen      36.7          19.3             193       3450 female 2007
6 Adelie   Torgersen      39.3          20.6             190       3650 male   2007
7 Adelie   Torgersen      38.9          17.8             181       3625 female 2007
8 Adelie   Torgersen      39.2          19.6             195       4675 male   2007
9 Adelie   Torgersen      34.1          18.1             193       3475 <NA>  2007
10 Adelie  Torgersen       42            20.2            190       4250 <NA>  2007
# ... with 334 more rows, and abbreviated variable names
#   1: flipper_length_mm,
#   2: body_mass_g
# i Use `print(n = ...)` to see more rows
```

For instance, the first penguin in the data is of the species Adelie (the three species included in the data are Adelie, Chinstrap and Gentoo), and was observed on the island called Torgeson. The remaining data for that penguin include measures of its bill length and depth, its flipper length and body mass, its sex and the year in which it was observed.

Note that though there are 344 rows in the tibble of data called `penguins`, only the first ten rows (`penguins`) are shown in the table above. Note also that the symbol `<NA>` is used to indicate a missing (not available) value.

2.3 Create newpenguins: Eliminating Missing Data

Next, let's take the `penguins` data from the `palmerpenguins` package, and identify those observations which have complete data (so, no missing values) in four variables of interest. We'll store that result in a new tibble (data set) called `new_penguins` and then take a look at that result using the following code.

Note that the code below:

- uses the “pipe” `|>` to send the penguins tibble to the `filter()` function
- uses `<-` to assign the result of our work to the `new_penguins` tibble
- uses the `complete.cases()` function to remove cases within `penguins` that have missing data on any of the four variables (`flipper_length_mm`, `body_mass_g`, `species` or `sex`) that we identify

```
new_penguins <- penguins |>
  filter(complete.cases(flipper_length_mm, body_mass_g, species, sex))

new_penguins

# A tibble: 333 x 8
  species island   bill_length_mm bill_depth_mm flipper_~1 body_~2 sex     year
  <fct>   <fct>        <dbl>        <dbl>       <int>    <int> <fct> <int>
1 Adelie  Torgersen      39.1       18.7       181     3750 male   2007
2 Adelie  Torgersen      39.5       17.4       186     3800 fema~  2007
3 Adelie  Torgersen      40.3        18         195     3250 fema~  2007
4 Adelie  Torgersen      36.7       19.3       193     3450 fema~  2007
5 Adelie  Torgersen      39.3       20.6       190     3650 male   2007
6 Adelie  Torgersen      38.9       17.8       181     3625 fema~  2007
7 Adelie  Torgersen      39.2       19.6       195     4675 male   2007
8 Adelie  Torgersen      41.1       17.6       182     3200 fema~  2007
9 Adelie  Torgersen      38.6       21.2       191     3800 male   2007
10 Adelie  Torgersen      34.6       21.1       198     4400 male   2007
# ... with 323 more rows, and abbreviated variable names 1: flipper_length_mm,
#   2: body_mass_g
# i Use `print(n = ...)` to see more rows
```

2.4 Counting Things and Making Tables

So, how many penguins are in our `new_penguins` data? When we printed out the result, we got an answer, but (as with many things in R) there are many ways to get the same result.

```
nrow(new_penguins)
```

```
[1] 333
```

How do our `new_penguins` data break down by sex and species? We'll use the `tabyl()` function from the `janitor` package to look at this.

```
new_penguins |>
  tabyl(sex, species)

  sex Adelie Chinstrap Gentoo
female    73      34      58
male     73      34      61
```

The output is reasonably clear (there are 73 female and 73 male Adelie penguins in the `newpenguins` tibble, for example) but could we make that table a little prettier, and while we're at it, can we add the row and column totals?

```
new_penguins |>
  tabyl(sex, species) |>
  adorn_totals(where = c("row", "col")) |> # add row, column totals
  kable() # one convenient way to make the table prettier
```

sex	Adelie	Chinstrap	Gentoo	Total
female	73	34	58	165
male	73	34	61	168
Total	146	68	119	333

The `kable()` function comes from the `knitr` package we loaded earlier. Notice that we added some comments to the code here with the prefix `#`. These comments are ignored by R in processing the data.

We can switch the rows and columns, and add some additional features, using the code below, which makes use of the `gt()` and `tab_header()` functions from the `gt` package, which is designed to help build complex tables. More on the incredibly versatile `gt()` package is available at <https://gt.rstudio.com/>.

```
new_penguins |>
  tabyl(species, sex) |>
  adorn_totals(where = c("row", "col")) |>
  gt() |>
  tab_header(
    title = md("Palmer Penguins in **newpenguins**"),
    subtitle = "Comparing sexes by species"
```

)

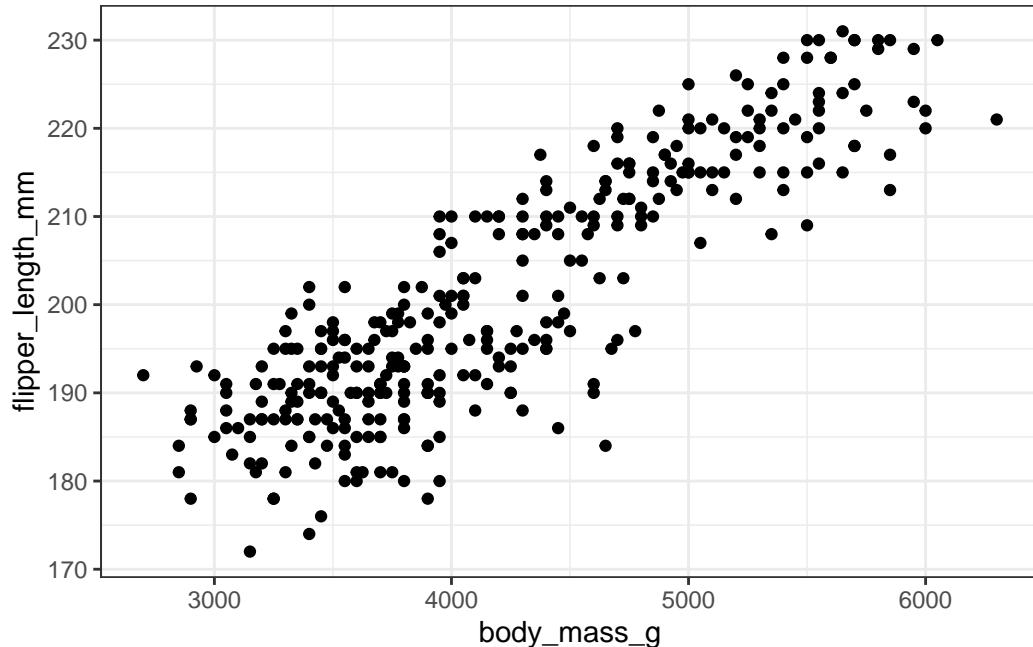
Palmer Penguins in `newpenguins` Comparing sexes by species

species	female	male	Total
Adelie	73	73	146
Chinstrap	34	34	68
Gentoo	58	61	119
Total	165	168	333

2.5 Creating a Scatterplot

Now, let's look at the other two variables of interest. Let's create a graph showing the association of body mass with flipper length across the complete set of 333 penguins.

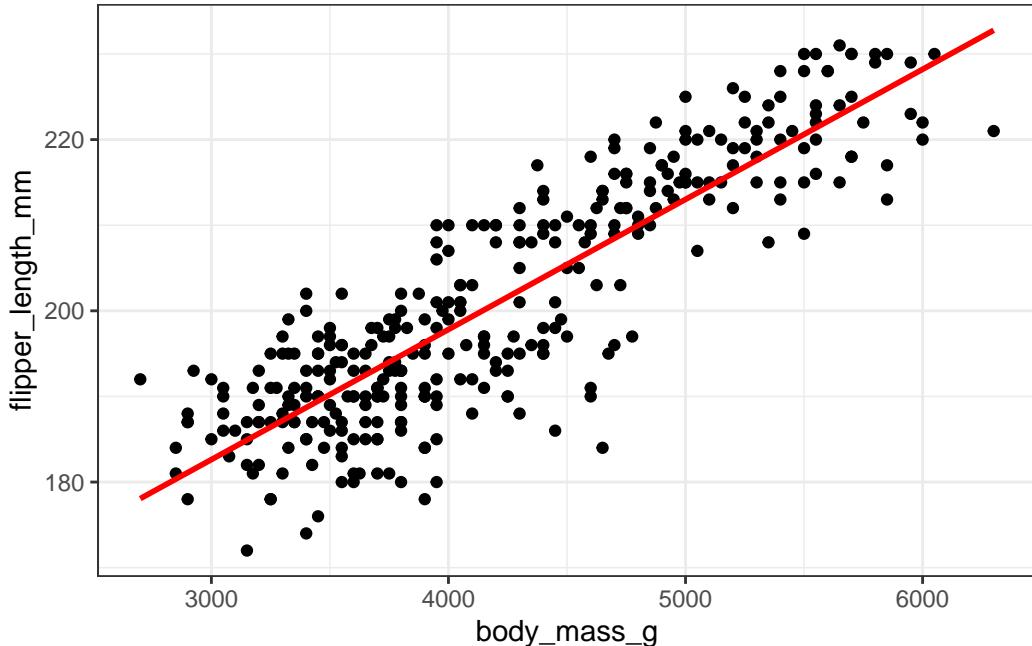
```
ggplot(new_penguins, aes(x = body_mass_g, y = flipper_length_mm)) +  
  geom_point()
```



Some of you may want to include a straight-line model (fit by a classical linear regression) to

this plot. One way to do that in R involves the addition of a single line of code, like this:

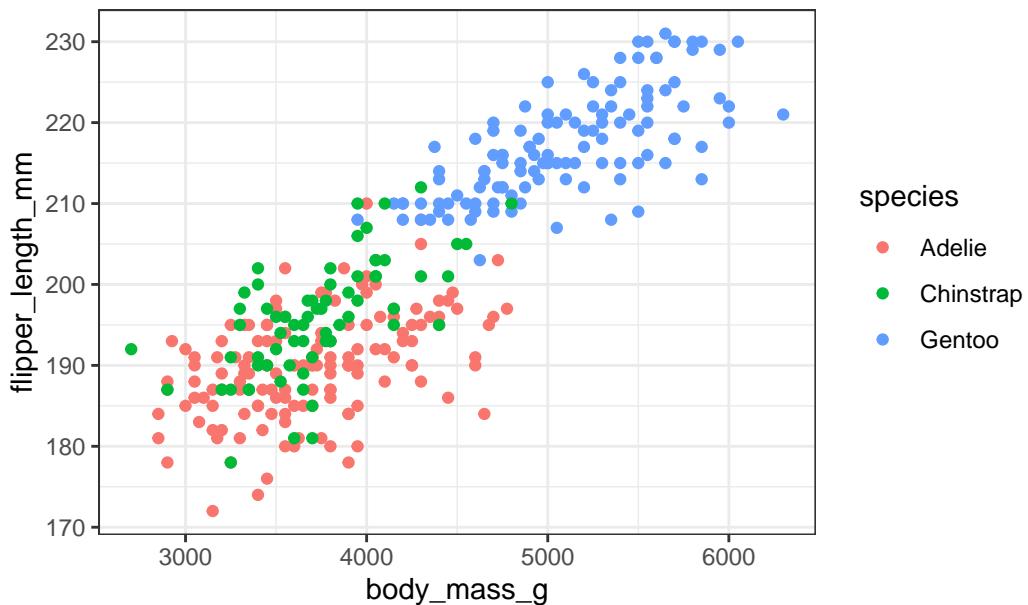
```
ggplot(new_penguins, aes(x = body_mass_g, y = flipper_length_mm)) +  
  geom_point() +  
  geom_smooth(method = "lm", formula = y ~ x,  
              col = "red", se = FALSE)
```



Whenever we build a graph for ourselves, these default choices may be sufficient. But I'd like to see a prettier version if I was going to show it to someone else. So, I might use a different color for each species, and I might add a title, like this.

```
ggplot(new_penguins, aes(x = body_mass_g, y = flipper_length_mm, col = species)) +  
  geom_point() +  
  labs(title = "Flipper Length and Body Mass for 333 of the Palmer Penguins")
```

Flipper Length and Body Mass for 333 of the Palmer Penguins



2.6 Six Ways To “Improve” This Graph

Now, let's build a new graph to incorporate some additional information and improve the appearance. Here, I want to:

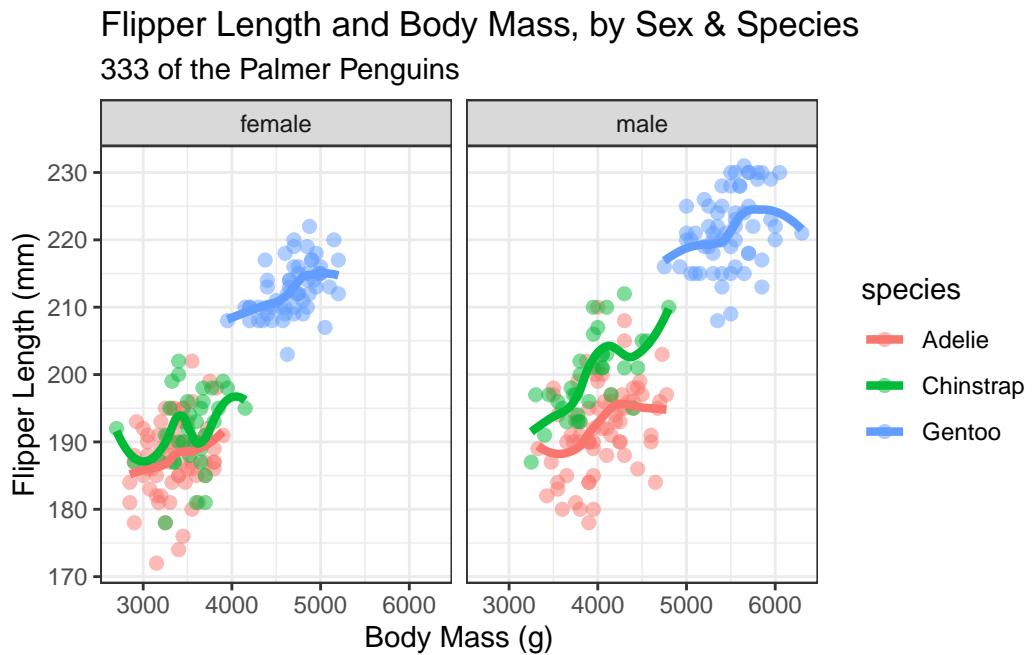
1. plot the relationship between body mass and flipper length in light of both Sex and Species
2. increase the size of the points and add a little transparency so we can see if points overlap,
3. add some smooth curves to summarize the relationships between the two quantities (body mass and flipper length) within each combination of species and sex,
4. split the graph into two “facets” (one for each sex),
5. improve the axis labels,
6. improve the titles by adding a subtitle, and also adding in some code to count the penguins (rather than hard-coding in the total number.)

```
ggplot(new_penguins, aes(x = body_mass_g, y = flipper_length_mm,
                           col = species)) +
  geom_point(size = 2, alpha = 0.5) +
  geom_smooth(method = "loess", formula = y ~ x,
              se = FALSE, size = 1.5) +
  facet_grid(~ sex) +
```

```

  labs(title = "Flipper Length and Body Mass, by Sex & Species",
       subtitle = str_glue(nrow(new_penguins), " of the Palmer Penguins"),
       x = "Body Mass (g)",
       y = "Flipper Length (mm)")

```



2.7 A Little Reflection

What can we learn from these plots and their construction? In particular,

- What do these plots suggest about the center of the distribution of each quantity (body mass and flipper length) overall, and within each combination of Sex and Species?
- What does the final plot suggest about the spread of the distribution of each of those quantities in each combination of Sex and Species?
- What do the plots suggest about the association of body mass and flipper length across the complete set of penguins?
- How does the shape and nature of this body mass - flipper length relationship change based on Sex and Species?
- Do you think it would be helpful to plot a straight-line relationship (rather than a smooth curve) within each combination of Sex and Species in the final plot? Why or why not? (Also, what would we have to do to the code to accomplish this?)

- How was the R code for the plot revised to accomplish each of the six “wants” specified above?

3 NHANES: A First Look

Next, we'll explore some data from the US National Health and Nutrition Examination Survey, or NHANES.

We'll display R code as we go, but we'll return to all of the key coding ideas involved later in the Notes.

3.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(NHANES)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

3.2 The NHANES data: A First Sample

The `NHANES` package provides a sample of 10,000 NHANES responses from the 2009-10 and 2011-12 administrations, in a data frame also called `NHANES`. We can obtain the dimensions of this data frame (think of it as a rectangle of data) with the `dim()` function.

```
dim(NHANES)
```

```
[1] 10000    76
```

We see that we have 10000 rows and 76 columns in the `NHANES` data frame.

For the moment, let's gather a random sample of 1,000 responses from the 10000 rows listed in the `NHANES` data frame, and then look at three variables (labeled Gender, Age and Height)

that describe those subjects¹. Some of the motivation for this example came from a Figure in Baumer, Kaplan, and Horton (2017).

```
# library(NHANES) # already loaded NHANES package/library of functions, data

set.seed(431001)
# use set.seed to ensure that we all get the same random sample
# of 1,000 NHANES subjects in our nh_1 collection

nh_1 <-
  slice_sample(NHANES, n = 1000, replace = FALSE) |>
  select(ID, SurveyYr, Gender, Age, Height)

nh_1

# A tibble: 1,000 x 5
  ID SurveyYr Gender   Age Height
  <int> <fct>   <fct> <int>  <dbl>
1 69638 2011_12 female    5   106.
2 70782 2011_12 male     64   176.
3 52408 2009_10 female   54   162.
4 59031 2009_10 female   15   155.
5 64530 2011_12 male    53   185.
6 71040 2011_12 male    63   169.
7 55186 2009_10 female  30   168.
8 60211 2009_10 male    5   103.
9 55730 2009_10 male    66   161.
10 68229 2011_12 female  36   170.
# ... with 990 more rows
# i Use `print(n = ...)` to see more rows
```

We have 1000 rows (observations) and 5 columns (variables) that describe the responses listed in the rows.

¹For more on the NHANES data available in the NHANES package, type ?NHANES in the Console in R Studio.

3.3 A Quick Numerical Summary

```
summary(nh_1)
```

ID	SurveyYr	Gender	Age	Height
Min. :51624	2009_10:512	female:504	Min. : 0.00	Min. : 85.0
1st Qu.:57011	2011_12:488	male :496	1st Qu.:18.00	1st Qu.:156.2
Median :61979			Median :36.00	Median :165.0
Mean :61903			Mean :37.42	Mean :162.3
3rd Qu.:67178			3rd Qu.:56.00	3rd Qu.:174.5
Max. :71875			Max. :80.00	Max. :195.9
				NA's :37

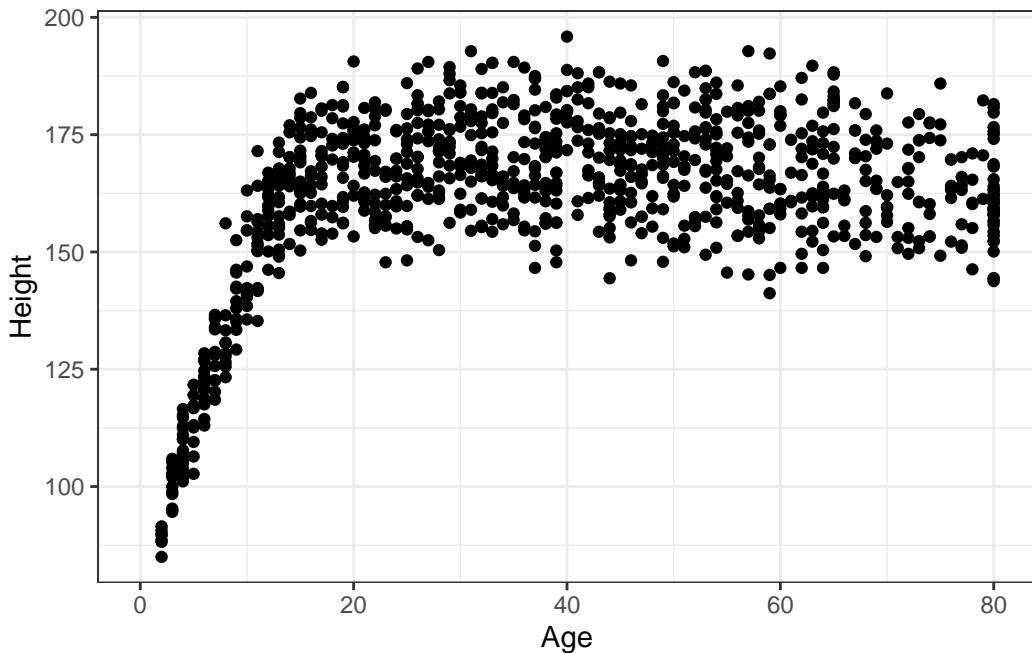
For the two variables that R recognizes as describing *categories*, SurveyYr and Gender, this numeric summary provides a small table of counts. For the Age and Height variables, we see the minimum, mean, maximum and other summary statistics.

3.4 Plotting Age vs. Height

Suppose we want to visualize the relationship of Height and Age in our 1,000 NHANES observations. The best choice is likely to be a scatterplot.

```
ggplot(data = nh_1, aes(x = Age, y = Height)) +  
  geom_point()
```

Warning: Removed 37 rows containing missing values (geom_point).



We note several interesting results here.

1. As a warning, R tells us that it has “Removed 37 rows containing missing values (geom_point).” Only 963 subjects plotted here, because the remaining 37 people have missing (NA) values for either Height, Age or both.
2. Unsurprisingly, the measured Heights of subjects grow from Age 0 to Age 20 or so, and we see that a typical Height increases rapidly across these Ages. The middle of the distribution at later Ages is pretty consistent at a Height somewhere between 150 and 175. The units aren’t specified, but we expect they must be centimeters. The Ages are clearly reported in Years.
3. No Age is reported over 80, and it appears that there is a large cluster of Ages at 80. This may be due to a requirement that Ages 80 and above be reported at 80 so as to help mask the identity of those individuals.²

As in this case, we’re going to build most of our visualizations using tools from the `ggplot2` package, which is part of the `tidyverse` series of packages. You’ll see similar coding structures throughout this Chapter, most of which are covered as well in Chapter 3 of Wickham and Gromelund (2022).

²If you visit the NHANES help file with `?NHANES`, you will see that subjects 80 years or older were indeed recorded as 80.

3.5 Restriction to Complete Cases

Before we move on, let's manipulate the data frame a bit, to focus on only those subjects who have complete data on both Age and Height. This will help us avoid that warning message.

```
nh_1cc <- nh_1 |>
  filter(complete.cases(Age, Height))

summary(nh_1cc)
```

ID	SurveyYr	Gender	Age	Height
Min. :51624	2009_10:487	female:484	Min. : 2.00	Min. : 85.0
1st Qu.:57034	2011_12:476	male :479	1st Qu.:19.00	1st Qu.:156.2
Median :62056			Median :37.00	Median :165.0
Mean :61967			Mean :38.29	Mean :162.3
3rd Qu.:67269			3rd Qu.:56.00	3rd Qu.:174.5
Max. :71875			Max. :80.00	Max. :195.9

Note that the units and explanations for these variables are contained in the NHANES help file, available via typing `?NHANES` in the Console of R Studio, or by typing `NHANES` into the Search bar in R Studio's Help window.

3.6 The Distinction between Gender and Sex

The `Gender` variable here is mis-named. These data refer to the biological status of these subjects, which is their `Sex`, and not the social construct of `Gender` which can be quite different. In our effort to avoid further confusion, we'll rename the variable `Gender` to `Sex` so as to more accurately describe what is actually measured here.

To do this, we can use this approach...

```
nh_1cc <- nh_1 |>
  rename(Sex = Gender) |>
  filter(complete.cases(Age, Height))

summary(nh_1cc)
```

ID	SurveyYr	Sex	Age	Height
Min. :51624	2009_10:487	female:484	Min. : 2.00	Min. : 85.0

```

1st Qu.:57034    2011_12:476    male   :479    1st Qu.:19.00    1st Qu.:156.2
Median  :62056                               Median :37.00    Median :165.0
Mean    :61967                               Mean  :38.29    Mean  :162.3
3rd Qu.:67269                               3rd Qu.:56.00    3rd Qu.:174.5
Max.    :71875                               Max.  :80.00    Max.  :195.9

```

That's better. How many observations do we have now? We could use `dim` to find out the number of rows and columns in this new data frame.

```
dim(nh_1cc)
```

```
[1] 963    5
```

Or, we could simply list the data frame and read off the result.

```
nh_1cc
```

```

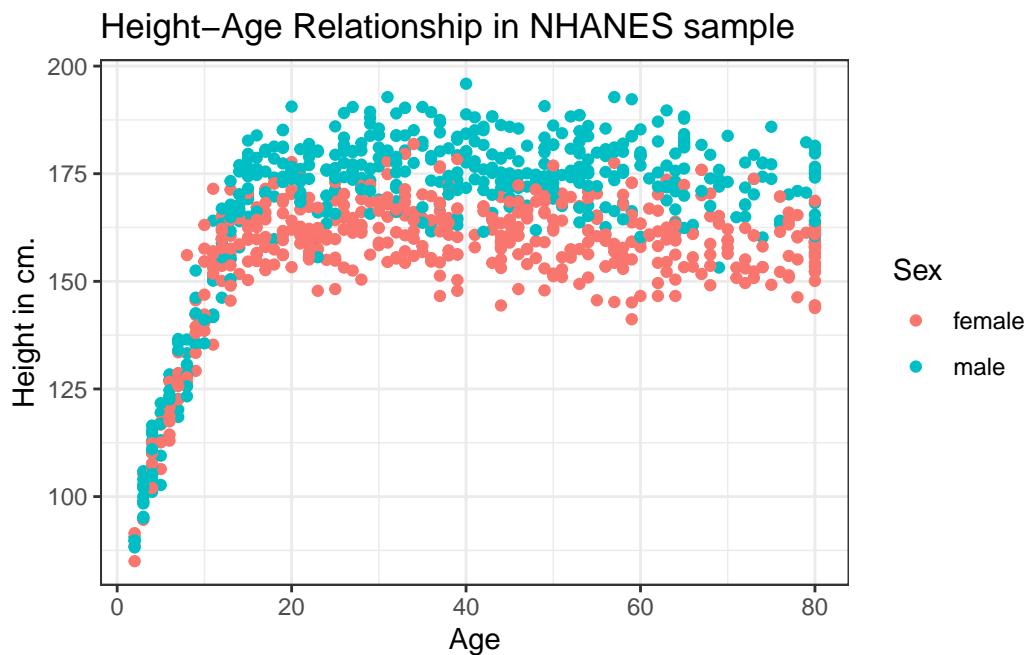
# A tibble: 963 x 5
  ID SurveyYr Sex     Age Height
  <int> <fct>   <fct> <int>  <dbl>
1 69638 2011_12 female   5    106.
2 70782 2011_12 male    64    176.
3 52408 2009_10 female   54    162.
4 59031 2009_10 female   15    155.
5 64530 2011_12 male    53    185.
6 71040 2011_12 male    63    169.
7 55186 2009_10 female   30    168.
8 60211 2009_10 male    5    103.
9 55730 2009_10 male    66    161.
10 68229 2011_12 female   36    170.
# ... with 953 more rows
# i Use `print(n = ...)` to see more rows

```

3.7 Age-Height by Sex?

Let's add Sex to the plot using color, and also adjust the y axis label to incorporate the units of measurement.

```
ggplot(data = nh_1cc, aes(x = Age, y = Height, color = Sex)) +
  geom_point() +
  labs(title = "Height-Age Relationship in NHANES sample",
       y = "Height in cm.")
```

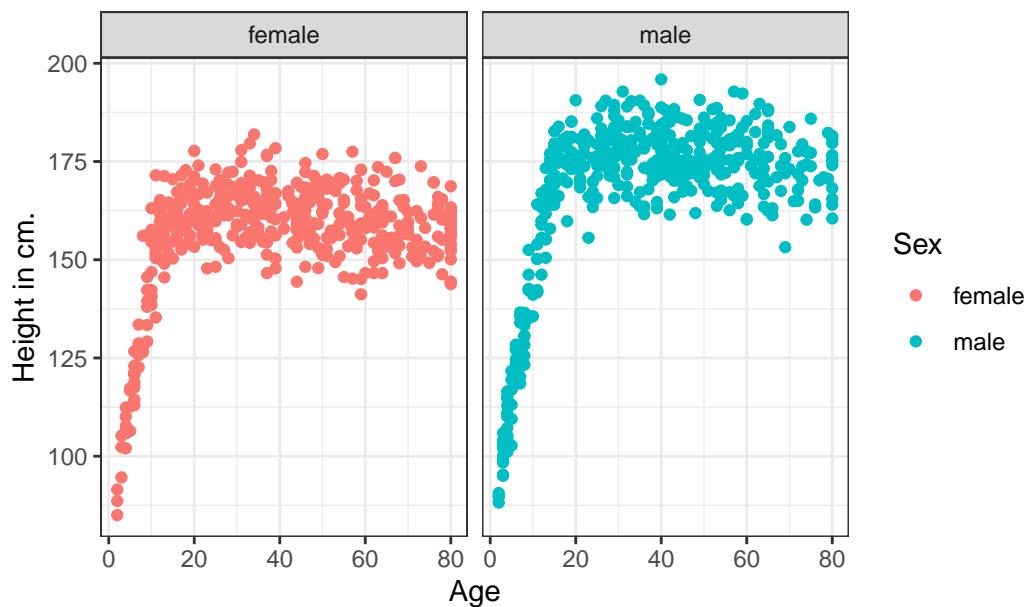


3.7.1 Can we show the Female and Male relationships in separate panels?

Sure.

```
ggplot(data = nh_1cc, aes(x = Age, y = Height, color = Sex)) +
  geom_point() +
  labs(title = "Height-Age Relationship in NHANES sample",
       y = "Height in cm.") +
  facet_wrap(~ Sex)
```

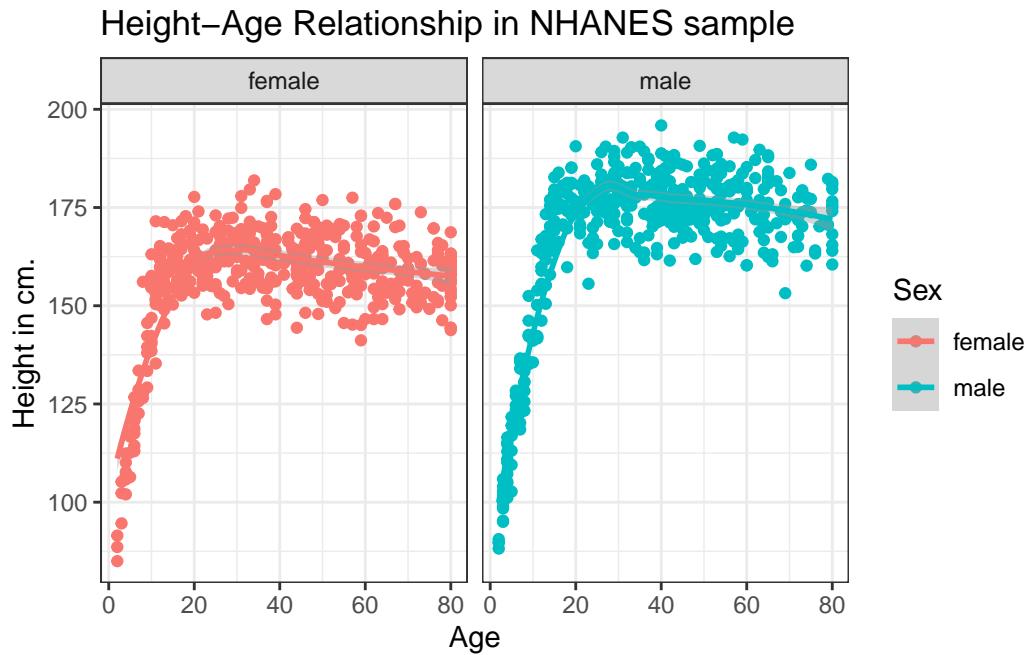
Height–Age Relationship in NHANES sample



3.7.2 Can we add a smooth curve to show the relationship in each plot?

Yes, by adding a call to the `geom_smooth()` function.

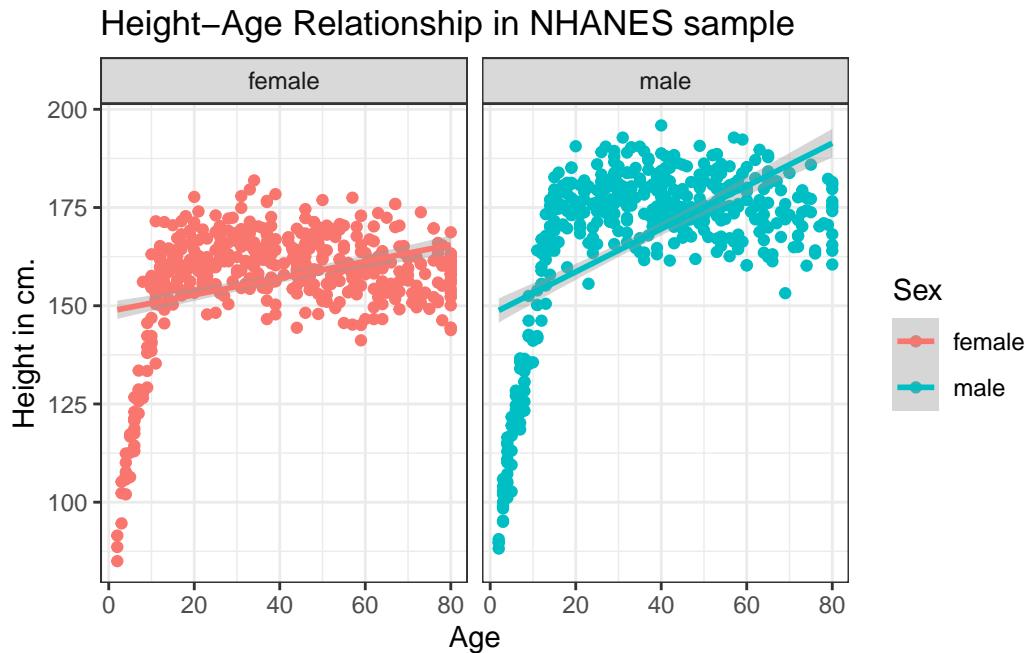
```
ggplot(data = nh_1cc, aes(x = Age, y = Height, color = Sex)) +  
  geom_point() +  
  geom_smooth(method = "loess", formula = y ~ x) +  
  labs(title = "Height–Age Relationship in NHANES sample",  
       y = "Height in cm.") +  
  facet_wrap(~ Sex)
```



3.7.3 What if we want to assume straight line relationships?

We could look at a linear model in each part of the plot instead. Does this make sense here?

```
ggplot(data = nh_1cc, aes(x = Age, y = Height, color = Sex)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Height–Age Relationship in NHANES sample",
       y = "Height in cm.") +
  facet_wrap(~ Sex)
```



It seems like the more complex relationship between Height and Age isn't well described by the straight line model.

3.8 Combining Plots with `patchwork`

The `patchwork` package in R allows us to use some simple commands to put two plots together.

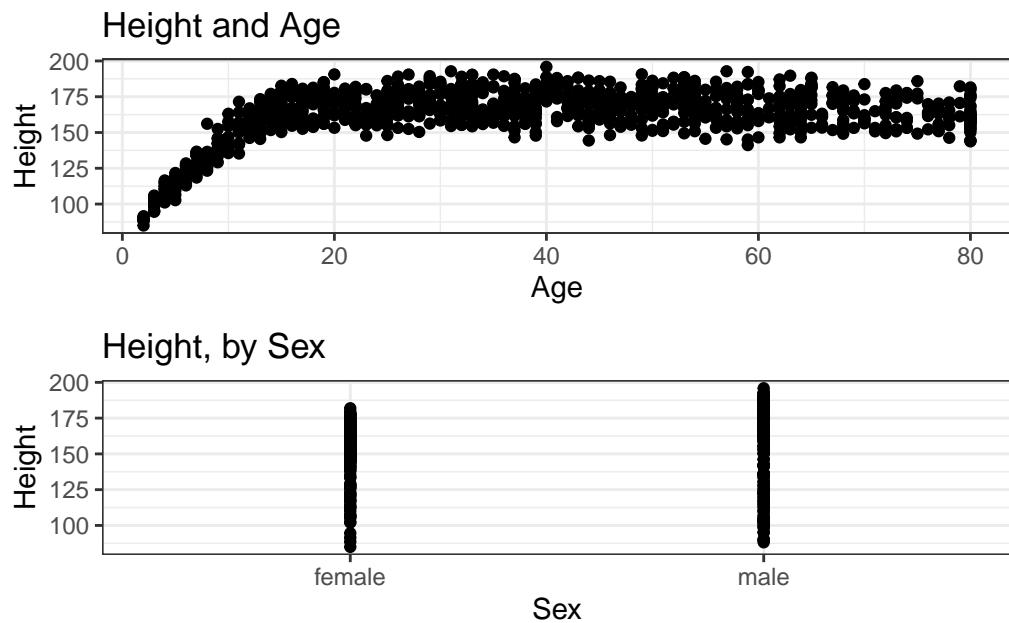
Suppose we create plots called `p1` and `p2`, as follows.

```
p1 <- ggplot(data = nh_1cc, aes(x = Age, y = Height)) +
  geom_point() +
  labs(title = "Height and Age")

p2 <- ggplot(data = nh_1cc, aes(x = Sex, y = Height)) +
  geom_point() +
  labs(title = "Height, by Sex")
```

Now, suppose we want to put them together in a single figure. Thanks to `patchwork`, we can simply type in the following.

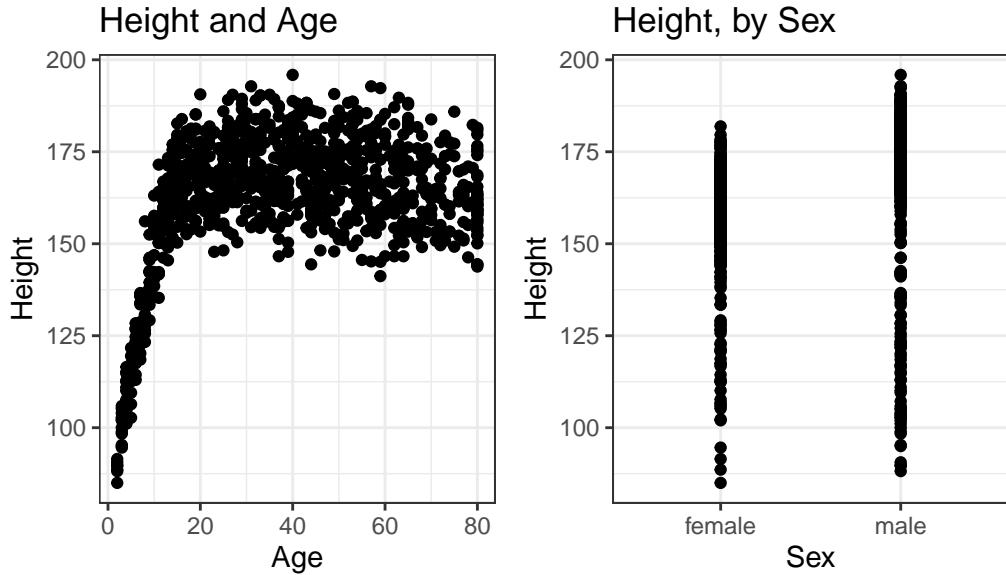
```
p1 / p2
```



or we can place the images next to each other, and add an annotation, like this:

```
p1 + p2 +
  plot_annotation(title = "Our Combined Plots")
```

Our Combined Plots



The [patchwork package website](#) provides lots of great examples and guides to make it very easy to combine separate ggplots into the same graphic. While there are other packages (`gridExtra` and `cowplot` are very nice, for instance) to do this task, I think `patchwork` is the most user-friendly, so that's the focus of these notes.

3.9 Coming Up

Next, we'll select a new sample of NHANES respondents a bit more carefully, introduce some new ways of thinking about data and variables, then we'll study those subjects in greater detail.

4 Data Structures, Variable Types & Sampling NHANES

4.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(NHANES)
library(janitor)
library(naniar)
library(tidyverse)

theme_set(theme_bw())
```

4.2 Data require structure and context

Descriptive statistics are concerned with the presentation, organization and summary of data, as suggested in Norman and Streiner (2014). This includes various methods of organizing and graphing data to get an idea of what those data can tell us.

As Vittinghoff et al. (2012) suggest, the nature of the measurement determines how best to describe it statistically, and the main distinction is between **numerical** and **categorical** variables. Even this is a little tricky - plenty of data can have values that look like numerical values, but are just numerals serving as labels.

As Bock, Velleman, and De Veaux (2004) point out, the truly critical notion, of course, is that data values, no matter what kind, are useless without their contexts. The Five W's (Who, What [and in what units], When, Where, Why, and often How) are just as useful for establishing the context of data as they are in journalism. If you can't answer Who and What, in particular, you don't have any useful information.

In general, each row of a data frame corresponds to an individual (respondent, experimental unit, record, or observation) about whom some characteristics are gathered in columns (and these characteristics may be called variables, factors or data elements.) Every column / variable

should have a name that indicates *what* it is measuring, and every row / observation should have a name that indicates *who* is being measured.

4.3 Sampling Adults in NHANES

In Chapter 3, we spent some time with a sample from the National Health and Nutrition Examination. Now, by changing the value of the `set.seed` function which determines the starting place for the random sampling, and changing some other specifications, we'll generate a new sample describing 750 unique (distinct) adult subjects who completed the 2011-12 version of the survey when they were between the ages of 21 and 64.

4.3.1 Creating a Temporary, Cleaner Data Frame

I'll start by describing the plan we will use to create a new data frame called `nh_temp` from which we will eventually build our final sample. In particular, let me lay out the steps I will use to create the `nh_temp` frame from the original NHANES data frame available in the R package called `NHANES`.

1. We'll **filter** the original NHANES data frame to include only the responses from the 2011-12 administration of the survey. This will cut the sample in half, from 10,000 rows to 5,000.
2. We'll then **filter** again to restrict the sample to adults whose age is at least 21 and also less than 65. I'll do this because I want to avoid problems with including both children and adults in my sample, and because I also want to focus on the population of people in the US who are usually covered by private insurance from their job, or by Medicaid insurance from the government, rather than those covered by Medicare.
3. As we discussed previously, what is listed in the NHANES data frame as `Gender` should be more correctly referred to as `Sex`. `Sex` is a biological feature of an individual, while `Gender` is a social construct. This is an important distinction, so I'll change the name of the variable.
4. We'll also rename three other variables, specifically we'll use `Race` to describe the `Race3` variable in the original NHANES data frame, as well as `SBP` to refer to the average systolic blood pressure, which is specified as `BPSysAve`, and `DBP` to refer to the average diastolic blood pressure, which is specified as `BPDiaAve`.
5. Having accomplished the previous four steps, we'll then **select** the variables we want to keep in the sample. (We use `select` for choosing variables or columns in the data frame, and `filter` for selecting subjects or rows.) The sixteen variables we will select are: `ID`, `Sex`, `Age`, `Height`, `Weight`, `Race`, `Education`, `BMI`, `SBP`, `DBP`, `Pulse`, `PhysActive`, `Smoke100`, `SleepTrouble`, `MaritalStatus` and `HealthGen`.

6. The original NHANES data frame includes some subjects (rows) multiple times in an effort to incorporate some of the sampling weights used in most NHANES analyses. For our purposes, though, we'd like to only include each subject one time. We use the `distinct()` function to limit the data frame to completely unique subjects (so that, for example, we don't wind up with two or more rows that have the same ID number.)

Here is the code I used to complete the six steps listed above and create the `nh_temp` data frame.

```
nh_temp <- NHANES |>
  filter(SurveyYr == "2011_12") |>
  filter(Age >= 21 & Age < 65) |>
  rename(Sex = Gender, Race = Race3, SBP = BPSysAve, DBP = BPDiaAve) |>
  select(ID, Sex, Age, Height, Weight, Race, Education, BMI, SBP, DBP,
         Pulse, PhysActive, Smoke100, SleepTrouble,
         MaritalStatus, HealthGen) |>
  distinct()
```

The resulting `nh_temp` data frame has 1700 rows and 16 columns.

```
nh_temp

# A tibble: 1,700 x 16
   ID Sex     Age Height Weight Race Educa~1   BMI    SBP    DBP Pulse PhysA~2
   <int> <fct> <int>  <dbl>  <dbl> <fct> <fct>   <dbl>  <int>  <int> <int> <fct>
 1 62172 fema~    43    172    98.6 Black High S~  33.3   103     72    80 No
 2 62176 fema~    34    172.   68.7 White Colleg~ 23.3   107     69    92 Yes
 3 62180 male     35    179.   89    White Colleg~ 27.9   107     66    66 No
 4 62199 male     57    186    96.9 White Colleg~ 28     110     65    84 Yes
 5 62205 male     28    171.   84.8 White Colleg~ 28.9   122     87    70 Yes
 6 62206 fema~    35    167.   81.5 White Some C~ 29.1   106     50    58 No
 7 62208 male     38    169.   63.2 Hisp~ Some C~ 22.2   105     59    52 Yes
 8 62209 fema~    62    143.   53.5 Mexi~ 8th Gr~ 26     108     57    72 No
 9 62220 fema~    31    167.   113. Black Colleg~ 40.4   120     71    62 Yes
10 62222 male     32    179    80.1 White Colleg~ 25     104     73    78 No
# ... with 1,690 more rows, 4 more variables: Smoke100 <fct>,
#   SleepTrouble <fct>, MaritalStatus <fct>, HealthGen <fct>, and abbreviated
#   variable names 1: Education, 2: PhysActive
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

4.3.2 Sampling nh_temp to obtain our nh_adult750 sample

Having established the `nh_temp` sampling frame, we now select a random sample of 750 adults from the 1700 available responses.

- We will use the `set.seed()` function in R to set a random numerical seed to ensure that if you redo this work, you will obtain the same sample.
 - Setting a seed is an important part of being able to replicate the work later when sampling is involved.
- Then we will use the `slice_sample()` function to actually draw the random sample, without replacement.
 - “Without replacement” means that once we’ve selected a particular subject, we won’t select them again.

```
set.seed(431002)
# use set.seed to ensure that we all get the same random sample

nh_adult750 <- slice_sample(nh_temp, n = 750, replace = F)

nh_adult750

# A tibble: 750 x 16
   ID Sex     Age Height Weight Race Educa~1   BMI    SBP    DBP Pulse PhysA~2
   <int> <fct> <int>  <dbl>  <dbl> <fct> <fct> <dbl> <int> <int> <int> <fct>
1 68648 fema~    30    181.   67.1 White Colleg~  20.4   103    59    78 No 
2 67200 male    30    180.   86.6 White Colleg~  26.7   113    68    70 Yes 
3 66404 fema~    35    160.   71.1 White Colleg~  27.8   116    80    68 Yes 
4 70535 male    40    177.    82    White Colleg~  26.3   130    79    68 No 
5 65308 fema~    54    151.   60.6 Mexi~ 8th Gr~  26.6   130    64    48 No 
6 67392 male    41    171.   90.7 Hisp~ Colleg~  31.2   124    82    68 Yes 
7 63218 male    35    163.    81    Mexi~ 8th Gr~  30.3   128    96    82 No 
8 65879 fema~    32    160.   66.4 Mexi~ Colleg~  25.9   104    70    78 Yes 
9 63617 male    29    189.   83.3 White Colleg~  23.2   105    72    76 Yes 
10 64720 male   29    174.   62.3 Black Colleg~  20.6   127    60    84 Yes
# ... with 740 more rows, 4 more variables: Smoke100 <fct>, SleepTrouble <fct>,
#   MaritalStatus <fct>, HealthGen <fct>, and abbreviated variable names
#   1: Education, 2: PhysActive
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

The `nh_adult750` data frame now includes 750 rows (observations) on 16 variables (columns). Essentially, we have 16 pieces of information on each of 750 adult NHANES subjects who were included in the 2011-12 panel.

4.3.3 Summarizing the Data's Structure

We can identify the number of rows and columns in a data frame or tibble with the `dim` function.

```
dim(nh_adult750)
```

```
[1] 750 16
```

The `str` function provides a lot of information about the structure of a data frame or tibble.

```
str(nh_adult750)
```

```
tibble [750 x 16] (S3: tbl_df/tbl/data.frame)
$ ID           : int [1:750] 68648 67200 66404 70535 65308 ...
$ Sex          : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 2 ...
$ Age          : int [1:750] 30 30 35 40 54 41 35 32 29 29 ...
$ Height       : num [1:750] 181 180 160 177 151 ...
$ Weight       : num [1:750] 67.1 86.6 71.1 82 60.6 90.7 81 66.4 83.3 62.3 ...
$ Race         : Factor w/ 6 levels "Asian","Black",...: 5 5 5 5 4 3 4 4 5 2 ...
$ Education    : Factor w/ 5 levels "8th Grade","9 - 11th Grade",...: 5 5 5 5 1 5 1 5 5 5 ...
$ BMI          : num [1:750] 20.4 26.7 27.8 26.3 26.6 31.2 30.3 25.9 23.2 20.6 ...
$ SBP          : int [1:750] 103 113 116 130 130 124 128 104 105 127 ...
$ DBP          : int [1:750] 59 68 80 79 64 82 96 70 72 60 ...
$ Pulse         : int [1:750] 78 70 68 68 48 68 82 78 76 84 ...
$ PhysActive   : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 2 1 2 2 2 ...
$ Smoke100     : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 1 2 1 2 2 ...
$ SleepTrouble : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 2 1 ...
$ MaritalStatus: Factor w/ 6 levels "Divorced","LivePartner",...: 3 4 3 3 2 3 3 3 3 2 ...
$ HealthGen    : Factor w/ 5 levels "Excellent","Vgood",...: 1 1 1 2 4 3 NA 1 2 4 ...
```

To see the first few observations, use `head`, and to see the last few, try `tail`...

```
tail(nh_adult750, 5) # shows the last five observations in the data set
```

```

# A tibble: 5 x 16
  ID Sex     Age Height Weight Race Educa~1   BMI    SBP    DBP Pulse PhysA~2
  <int> <fct> <int>  <dbl>  <dbl> <fct> <fct>   <dbl> <int> <int> <int> <fct>
1 63924 female    29    165.   113. Black High S~  41.9    98     56    74 No
2 69825 female    43    164.   63.3 White Colleg~ 23.7   122     83    88 Yes
3 68109 male      45    170.   78.7 Black High S~  27.1   140     79   102 Yes
4 64598 female    60    158.   74.5 White Some C~  29.8   137     80    78 Yes
5 64048 female    54    161.   67.5 White Some C~  26.2   121     87    72 No
# ... with 4 more variables: Smoke100 <fct>, SleepTrouble <fct>,
#   MaritalStatus <fct>, HealthGen <fct>, and abbreviated variable names
#   1: Education, 2: PhysActive
# i Use `colnames()` to see all variable names

```

4.3.4 What are the variables?

We can use the `glimpse` function to get a short preview of the data.

```
glimpse(nh_adult750)
```

```

Rows: 750
Columns: 16
$ ID           <int> 68648, 67200, 66404, 70535, 65308, 67392, 63218, 65879, ~
$ Sex          <fct> female, male, female, male, female, male, male, female, ~
$ Age          <int> 30, 30, 35, 40, 54, 41, 35, 32, 29, 29, 64, 28, 31, 59, ~
$ Height        <dbl> 181.3, 180.2, 159.8, 176.6, 150.9, 170.6, 163.4, 160.2, ~
$ Weight        <dbl> 67.1, 86.6, 71.1, 82.0, 60.6, 90.7, 81.0, 66.4, 83.3, 62-
$ Race          <fct> White, White, White, White, Mexican, Hispanic, Mexican, ~
$ Education     <fct> College Grad, College Grad, College Grad, College Grad, ~
$ BMI           <dbl> 20.4, 26.7, 27.8, 26.3, 26.6, 31.2, 30.3, 25.9, 23.2, 20-
$ SBP            <int> 103, 113, 116, 130, 130, 124, 128, 104, 105, 127, 128, 1-
$ DBP            <int> 59, 68, 80, 79, 64, 82, 96, 70, 72, 60, 74, 76, 82, 66, ~
$ Pulse          <int> 78, 70, 68, 68, 48, 68, 82, 78, 76, 84, 62, 56, 78, 66, ~
$ PhysActive     <fct> No, Yes, Yes, No, No, Yes, Yes, Yes, Yes, Yes, No, N-
$ Smoke100       <fct> No, Yes, No, Yes, Yes, No, Yes, Yes, Yes, No, No, Ye-
$ SleepTrouble   <fct> Yes, No, No, No, No, No, Yes, No, No, Yes, Yes, No, Y-
$ MaritalStatus  <fct> Married, NeverMarried, Married, Married, LivePartner, Ma-
$ HealthGen      <fct> Excellent, Excellent, Excellent, Vgood, Fair, Good, NA, ~

```

The variables we have collected are described in the brief table below¹.

¹Descriptions are adapted from the ?NHANES help file. Remember that what NHANES lists as Gender is

Variable	Description	Sample Values
ID	a numerical code identifying the subject	68648, 67200
Sex	sex of subject (2 levels)	female, male
Age	age (years) at screening of subject	30, 35
Height	height (in cm) at screening of subject	181.3, 180.2
Weight	weight (in kg) at screening of subject	67.1, 86.6
Race	reported race of subject (6 levels)	White, Black
Education	educational level of subject (5 levels)	College Grad, High School
BMI	body-mass index, in kg/m ²	20.4, 26.7
SBP	systolic blood pressure in mm Hg	103, 113
DBP	diastolic blood pressure in mm Hg	59, 68
Pulse	60 second pulse rate in beats per minute	78, 70
PhysActive	Moderate or vigorous-intensity sports?	Yes, No
Smoke100	Smoked at least 100 cigarettes lifetime?	Yes, No
SleepTrouble	Told a doctor they have trouble sleeping?	Yes, No
MaritalStatus	Marital Status	Married, Divorced
HealthGen	Self-report general health rating (5 levels)	Vgood, Fair

The levels for the multi-categorical variables are:

- **Race:** Mexican, Hispanic, White, Black, Asian, or Other.
- **Education:** 8th Grade, 9 - 11th Grade, High School, Some College, or College Grad.
- **MaritalStatus:** Married, Widowed, Divorced, Separated, NeverMarried or LivePartner (living with partner).
- **HealthGen:** Excellent, Vgood, Good, Fair or Poor.

Some details can be obtained using the **summary** function.

```
summary(nh_adult750)
```

ID	Sex	Age	Height	Weight
Min. :62206	female:388	Min. :21.00	Min. :142.4	Min. : 39.30
1st Qu.:64277	male :362	1st Qu.:30.00	1st Qu.:161.8	1st Qu.: 67.40
Median :66925		Median :40.00	Median :168.9	Median : 80.00
Mean :66936		Mean :40.82	Mean :168.9	Mean : 83.16
3rd Qu.:69414		3rd Qu.:51.00	3rd Qu.:175.7	3rd Qu.: 95.30
Max. :71911		Max. :64.00	Max. :200.4	Max. :198.70

captured here as Sex, and similarly Race3, BPSysAve and BPDiaAve from NHANES are here listed as Race, SBP and DBP.

Race	Education	BMI	NA's :5	NA's :5	SBP
Asian : 70	8th Grade : 50	Min. :16.70	Min. : 83.0		
Black :128	9 - 11th Grade: 76	1st Qu.:24.20	1st Qu.:108.0		
Hispanic: 63	High School :143	Median :27.90	Median :118.0		
Mexican : 80	Some College :241	Mean :29.08	Mean :118.8		
White :393	College Grad :240	3rd Qu.:32.10	3rd Qu.:127.0		
Other : 16		Max. :80.60	Max. :209.0		
		NA's :5	NA's :33		
DBP	Pulse	PhysActive	Smoke100	SleepTrouble	
Min. : 0.00	Min. : 40.00	No :326	No :453	No :555	
1st Qu.: 66.00	1st Qu.: 66.00	Yes:424	Yes:297	Yes:195	
Median : 73.00	Median : 72.00				
Mean : 72.69	Mean : 73.53				
3rd Qu.: 80.00	3rd Qu.: 80.00				
Max. :108.00	Max. :124.00				
NA's :33	NA's :32				
MaritalStatus	HealthGen				
Divorced : 78	Excellent: 84				
LivePartner : 70	Vgood :197				
Married :388	Good :252				
NeverMarried:179	Fair :104				
Separated : 19	Poor : 14				
Widowed : 16	NA's : 99				

Note the appearance of NA's (indicating missing values) in some columns, and that some variables are summarized by a list of their (categorical) values (with counts) and some (quantitative/numeric) variables are summarized with a minimum, quartiles and means.

4.4 Quantitative Variables

Variables recorded in numbers that we use as numbers are called **quantitative**. Familiar examples include incomes, heights, weights, ages, distances, times, and counts. All quantitative variables have measurement units, which tell you how the quantitative variable was measured. Without units (like miles per hour, angstroms, yen or degrees Celsius) the values of a quantitative variable have no meaning.

- It does little good to be told the price of something if you don't know the currency being used.

- You might be surprised to see someone whose age is 72 listed in a database on childhood diseases until you find out that age is measured in months.
- Often just seeking the units can reveal a variable whose definition is challenging - just how do we measure “friendliness”, or “success,” for example.
- Quantitative variables may also be classified by whether they are **continuous** or can only take on a **discrete** set of values. Continuous data may take on any value, within a defined range. Suppose we are measuring height. While height is really continuous, our measuring stick usually only lets us measure with a certain degree of precision. If our measurements are only trustworthy to the nearest centimeter with the ruler we have, we might describe them as discrete measures. But we could always get a more precise ruler. The measurement divisions we make in moving from a continuous concept to a discrete measurement are usually fairly arbitrary. Another way to think of this, if you enjoy music, is that, as suggested in Norman and Streiner (2014), a piano is a *discrete* instrument, but a violin is a *continuous* one, enabling finer distinctions between notes than the piano is capable of making. Sometimes the distinction between continuous and discrete is important, but usually, it's not.
 - The `nh_adult750` data includes several quantitative variables, specifically `Age`, `Height`, `BMI`, `SBP`, `DBP` and `Pulse`.
 - We know these are quantitative because they have units: `Age` in years, `Height` in centimeters, `BMI` in kg/m^2 , the `BP` measurements in mm Hg, and `Pulse` in beats per minute.
 - Depending on the context, we would likely treat most of these as *discrete* given that the measurements are fairly crude (this is certainly true for `Age`, measured in years) although `BMI` is probably *continuous* in most settings, even though it is a function of two other measures (`Height` and `Weight`) which are rounded off to integer numbers of centimeters and kilograms, respectively.
- It is also possible to separate out quantitative variables into **ratio** variables or **interval** variables. An interval variable has equal distances between values, but the zero point is arbitrary. A ratio variable has equal intervals between values, and a meaningful zero point. For example, weight is an example of a ratio variable, while IQ is an example of an interval variable. We all know what zero weight is. An intelligence score like IQ is a different matter. We say that the average IQ is 100, but that's only by convention. We could just as easily have decided to add 400 to every IQ value and make the average 500 instead. Because IQ's intervals are equal, the difference between an IQ of 70 and an IQ of 80 is the same as the difference between 120 and 130. However, an IQ of 100 is not twice as high as an IQ of 50. The point is that if the zero point is artificial and movable, then the differences between numbers are meaningful but the ratios between them are not. On the other hand, most lab test values are ratio variables, as are physical characteristics like height and weight. A person who weighs 100 kg is twice as heavy as

one who weighs 50 kg; even when we convert kg to pounds, this is still true. For the most part, we can treat and analyze interval or ratio variables the same way.

- Each of the quantitative variables in our `nh_adult750` data can be thought of as ratio variables.
- Quantitative variables lend themselves to many of the summaries we will discuss, like means, quantiles, and our various measures of spread, like the standard deviation or inter-quartile range. They also have at least a chance to follow the Normal distribution.

4.4.1 A look at BMI (Body-Mass Index)

The definition of BMI (*body-mass index*) for adult subjects (which is expressed in units of kg/m^2) is:

$$\text{Body Mass Index} = \frac{\text{weight in kg}}{(\text{height in meters})^2} = 703 \times \frac{\text{weight in pounds}}{(\text{height in inches})^2}$$

[BMI is essentially] ... a measure of a person's *thinness* or *thickness*... BMI was designed for use as a simple means of classifying average sedentary (physically inactive) populations, with an average body composition. For these individuals, the current value recommendations are as follows: a BMI from 18.5 up to 25 may indicate optimal weight, a BMI lower than 18.5 suggests the person is underweight, a number from 25 up to 30 may indicate the person is overweight, and a number from 30 upwards suggests the person is obese.

Wikipedia, https://en.wikipedia.org/wiki/Body_mass_index

4.5 Qualitative (Categorical) Variables

Qualitative or categorical variables consist of names of categories. These names may be numerical, but the numbers (or names) are simply codes to identify the groups or categories into which the individuals are divided. Categorical variables with two categories, like yes or no, up or down, or, more generally, 1 and 0, are called **binary** variables. Those with more than two-categories are sometimes called **multi-categorical** variables.

- When the categories included in a variable are merely names, and come in no particular order, we sometimes call them **nominal** variables. The most important summary of such a variable is usually a table of frequencies, and the mode becomes an important single summary, while the mean and median are essentially useless.

- In the `nh_adult750` data, `Race` is a nominal variable with multiple unordered categories. So is `MaritalStatus`.
- The alternative categorical variable (where order matters) is called **ordinal**, and includes variables that are sometimes thought of as falling right in between quantitative and qualitative variables.
 - Examples of ordinal multi-categorical variables in the `nh_adult750` data include the `Education` and `HealthGen` variables.
 - Answers to questions like “How is your overall physical health?” with available responses Excellent, Very Good, Good, Fair or Poor, which are often coded as 1-5, certainly provide a perceived *order*, but a group of people with average health status 4 (Very Good) is not necessarily twice as healthy as a group with average health status of 2 (Fair).
- Sometimes we treat the values from ordinal variables as sufficiently scaled to permit us to use quantitative approaches like means, quantiles, and standard deviations to summarize and model the results, and at other times, we’ll treat ordinal variables as if they were nominal, with tables and percentages our primary tools.
- Note that all binary variables may be treated as ordinal, or nominal.
 - Binary variables in the `nh_adult750` data include `Sex`, `PhysActive`, `Smoke100`, `SleepTrouble`. Each can be thought of as either ordinal or nominal.

Lots of variables may be treated as either quantitative or qualitative, depending on how we use them. For instance, we usually think of age as a quantitative variable, but if we simply use age to make the distinction between “child” and “adult” then we are using it to describe categorical information. Just because your variable’s values are numbers, don’t assume that the information provided is quantitative.

4.6 Counting Missing Values

The `summary()` command counts the number of missing observations in each variable, but sometimes you want considerably more information.

We can use some functions from the `naniar` package to learn useful things about the missing data in our `nh_adult750` sample.

The `miss_var_table` command provides a table of the number of variables with 0, 1, 2, up to n, missing values and the percentage of the total number of variables those variables make up.

```
miss_var_table(nh_adult750)
```

```
# A tibble: 5 x 3
  n_miss_in_var n_vars pct_vars
  <int>    <int>    <dbl>
1       0        9     56.2
2       5        3     18.8
3      32        1      6.25
4      33        2     12.5
5      99        1      6.25
```

So, for instance, we have 9 variables with no missing data, and that constitutes 56.25% of the 16 variables in our `nh_adult750` data.

The `miss_var_summary()` function tabulates the number, percent missing, and cumulative sum of missing of each variable in our data frame, in order of most to least missing values.

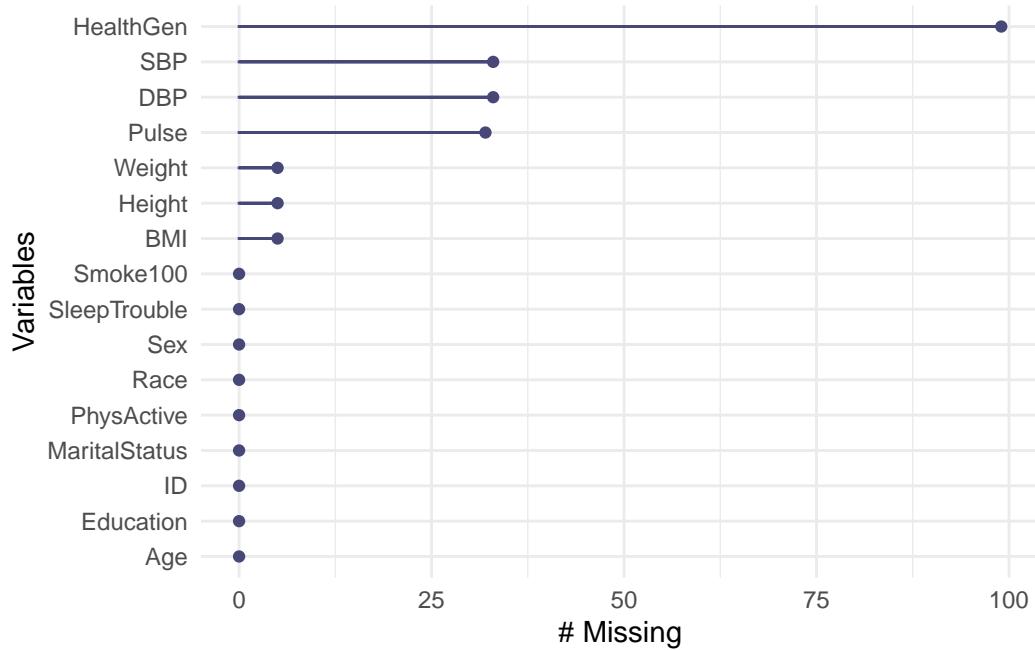
```
miss_var_summary(nh_adult750)
```

```
# A tibble: 16 x 3
  variable      n_miss pct_miss
  <chr>        <int>    <dbl>
1 HealthGen      99     13.2
2 SBP            33      4.4
3 DBP            33      4.4
4 Pulse          32      4.27
5 Height          5     0.667
6 Weight          5     0.667
7 BMI             5     0.667
8 ID              0      0
9 Sex              0      0
10 Age             0      0
11 Race            0      0
12 Education       0      0
13 PhysActive      0      0
14 Smoke100        0      0
15 SleepTrouble     0      0
16 MaritalStatus     0      0
```

So, for example, the `rmiss_var_summary(nh_temp) |> slice_head(n = 1) |> select(variable)` variable is the one missing more of our data than anything else within the `nh_adult750`` data frame.

A graph of this information is available, as well.

```
gg_miss_var(nh_adult750)
```



I'll note that there are also functions to count the number of missing observations by case (observation) rather than variable. For example, we can use `miss_case_table`.

```
miss_case_table(nh_adult750)
```

```
# A tibble: 6 x 3
  n_miss_in_case n_cases pct_cases
    <int>     <int>     <dbl>
1         0       636   84.8
2         1        78   10.4
3         3        15     2
4         4        19   2.53
5         6         1  0.133
6         7         1  0.133
```

Now we see that 636 observations, or 84.8% of all cases have no missing data.

We can use `miss_case_summary()` to identify cases with missing data, as well.

```

miss_case_summary(nh_adult750)

# A tibble: 750 x 3
  case n_miss pct_miss
  <int>   <int>    <dbl>
1 342      7     43.8
2 606      6     37.5
3 157      4      25
4 169      4      25
5 204      4      25
6 234      4      25
7 323      4      25
8 415      4      25
9 478      4      25
10 483     4      25
# ... with 740 more rows
# i Use `print(n = ...)` to see more rows

```

4.7 nh_adults500cc: A Sample of Complete Cases

If we wanted a sample of exactly 750 subjects with complete data, we would have needed to add a step in the development of our `nh_temp` sampling frame to filter for complete cases.

```

nh_temp2 <- NHANES |>
  filter(SurveyYr == "2011_12") |>
  filter(Age >= 21 & Age < 65) |>
  rename(Sex = Gender, Race = Race3, SBP = BPSysAve, DBP = BPDiaAve) |>
  select(ID, Sex, Age, Height, Weight, Race, Education, BMI, SBP, DBP,
         Pulse, PhysActive, Smoke100, SleepTrouble,
         MaritalStatus, HealthGen) |>
  distinct() |>
  na.omit()

```

Let's check that this new sampling frame has no missing data.

```

miss_var_table(nh_temp2)

# A tibble: 1 x 3
n_miss_in_var n_vars pct_vars

```

```

<int> <int> <dbl>
1       0     16    100

```

OK. Now, let's create a second sample, called `nh_adult500cc`, where now, we will select 500 adults with complete data on all of the variables of interest, and using a different random seed. The `cc` here stands for complete cases.

```

set.seed(431003)
# use set.seed to ensure that we all get the same random sample

nh_adult500cc <- slice_sample(nh_temp2, n = 500, replace = F)

nh_adult500cc

# A tibble: 500 x 16
   ID Sex     Age Height Weight Race Educa~1   BMI    SBP    DBP Pulse PhysA~2
   <int> <fct> <int> <dbl> <dbl> <fct> <fct> <dbl> <int> <int> <int> <fct>
1 64079 fema~    25    159.   86.2 Hisp~ Some C~  34.2   120    67    84 Yes
2 64374 fema~    52    169    65.5 Asian Colleg~ 22.9    92    58    60 Yes
3 71875 male     42    182.   94.1 Black Colleg~ 28.5   102    63    76 Yes
4 66396 fema~    46    161.   107.  Asian 8th Gr~ 41.2   111    61    70 No
5 64315 fema~    52    161.   64.5 White 9 - 11~ 24.9   130    69    68 Yes
6 64015 male     32    168.   82.3 Mexi~ Some C~  29     119    79    70 No
7 63590 male     21    181.   98.3 Black Some C~ 29.9   121    67    58 Yes
8 70893 fema~    30    171.   65.7 White 9 - 11~ 22.5   104    75    74 Yes
9 70828 male     26    178.   100.  White Some C~ 31.5   119    77    66 No
10 67930 male    59    172.   91.7 Mexi~ Colleg~ 31     127    85    66 No
# ... with 490 more rows, 4 more variables: Smoke100 <fct>, SleepTrouble <fct>,
#   MaritalStatus <fct>, HealthGen <fct>, and abbreviated variable names
#   1: Education, 2: PhysActive
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names

```

4.8 Saving our Samples in .Rds files

We'll save the `nh_adult750` and `nh_adult500cc` samples to use in later parts of the notes. To do this, we'll save them as `.Rds` files, which will have some advantages for us later on.

```

write_rds(nh_adult750, file = "data/nh_adult750.Rds")
write_rds(nh_adult500cc, file = "data/nh_adult500cc.Rds")

```

You will also find these `.Rds` files as part of the [431-data repository](#) for the course. Next, we'll load, explore and learn about some of the variables in these two samples.

5 Visualizing NHANES Data

5.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(janitor)
library(knitr) ## for kable
library(tidyverse)

theme_set(theme_bw())
```

5.2 Loading in the “Complete Cases” Sample

Let’s begin by loading into the `nh_500cc` data frame the information from the `nh_adult500cc.Rds` file we created in Section @ref(nh_cc).

```
nh_500cc <- read_rds("data/nh_adult500cc.Rds")
```

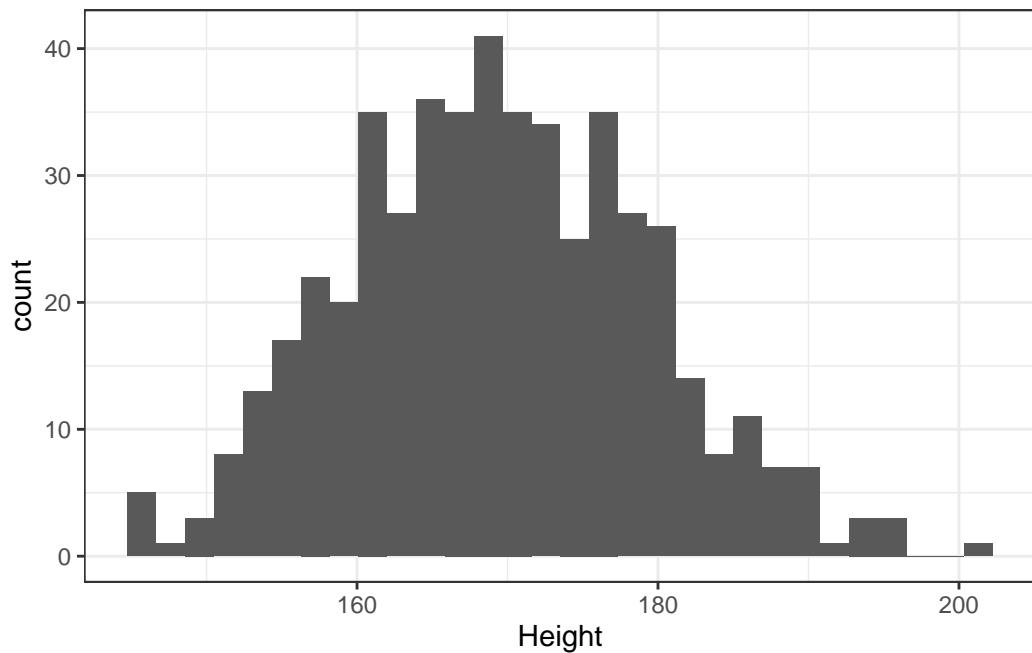
One obvious hurdle we’ll avoid for the moment is what to do about missing data, since the `nh_500cc` data are specifically drawn from complete responses. Working with complete cases only can introduce bias to our estimates and visualizations, so it will be necessary in time to address what we should do when a complete-case analysis isn’t a good choice. We’ll return to this issue in a few chapters.

5.3 Distribution of Heights

What is the distribution of height in this new sample?

```
ggplot(data = nh_500cc, aes(x = Height)) +
  geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

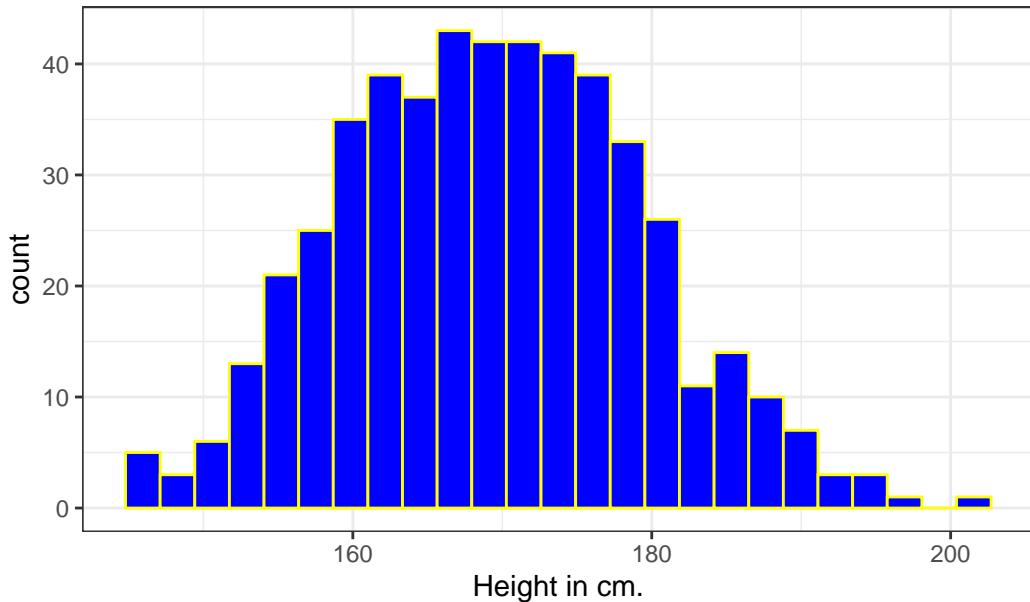


We can do several things to clean this up.

1. We'll change the color of the lines for each bar of the histogram.
2. We'll change the fill inside each bar to make them stand out a bit more.
3. We'll add a title and relabel the horizontal (x) axis to include the units of measurement.
4. We'll avoid the warning by selecting a number of bins (we'll use 25 here) into which we'll group the heights before drawing the histogram.

```
ggplot(data = nh_500cc, aes(x = Height)) +  
  geom_histogram(bins = 25, col = "yellow", fill = "blue") +  
  labs(title = "Height of NHANES subjects ages 21-64",  
       x = "Height in cm.")
```

Height of NHANES subjects ages 21–64

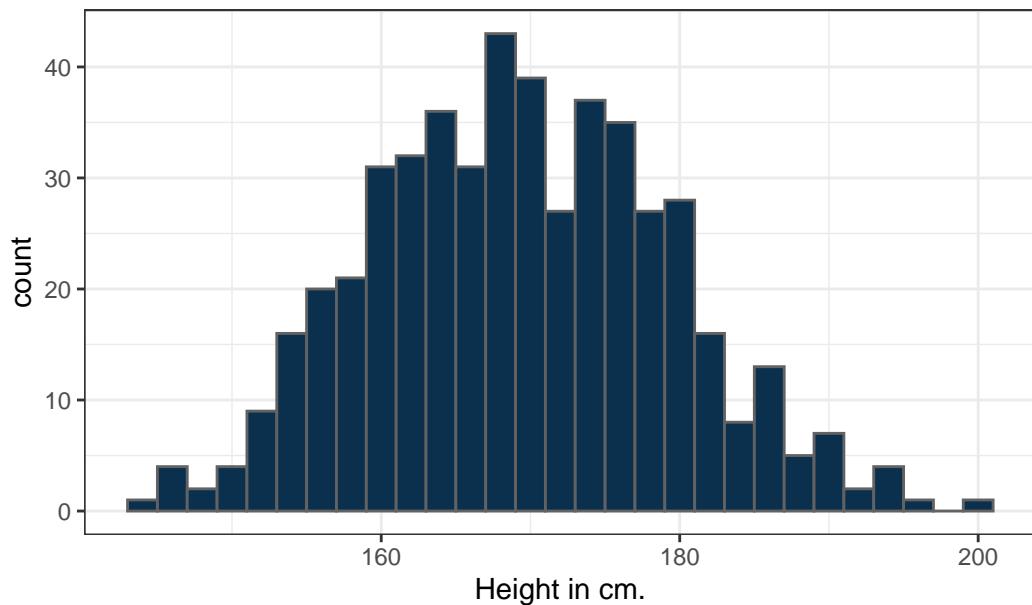


5.3.1 Changing a Histogram's Fill and Color

The CWRU color guide (<https://case.edu/umc/our-brand/visual-guidelines/>) lists the HTML color schemes for CWRU blue and CWRU gray. Let's match that color scheme. We will also change the bins for the histogram, to gather observations into groups of 2 cm. each, by specifying the width of the bins, rather than the number of bins.

```
cwru.blue <- '#0a304e'  
cwru.gray <- '#626262'  
  
ggplot(data = nh_500cc, aes(x = Height)) +  
  geom_histogram(binwidth = 2,  
                 col = cwru.gray, fill = cwru.blue) +  
  labs(title = "Height of NHANES subjects ages 21–64",  
        x = "Height in cm.")
```

Height of NHANES subjects ages 21–64

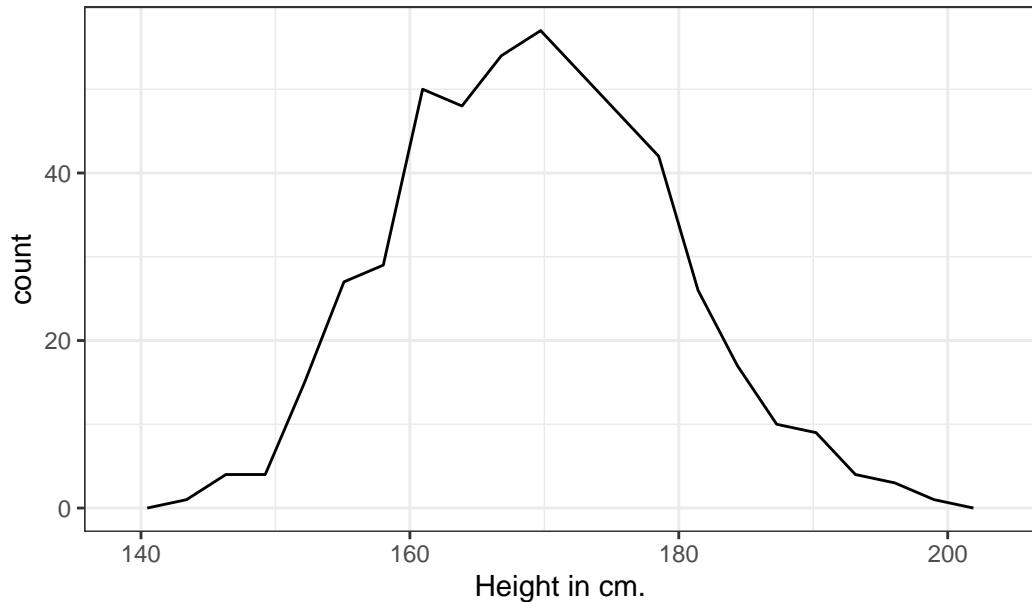


5.3.2 Using a frequency polygon

A frequency polygon essentially smooths out the top of the histogram, and can also be used to show the distribution of Height.

```
ggplot(data = nh_500cc, aes(x = Height)) +  
  geom_freqpoly(bins = 20) +  
  labs(title = "Height of NHANES subjects ages 21-64",  
       x = "Height in cm.")
```

Height of NHANES subjects ages 21–64

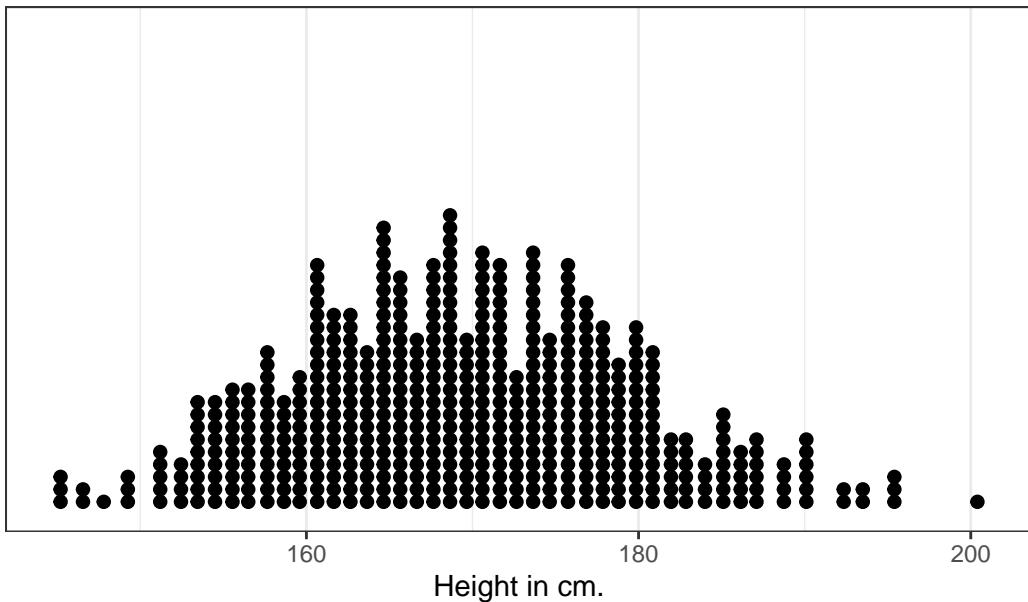


5.3.3 Using a dotplot

A dotplot can also be used to show the distribution of a variable like `Height`, and produces a somewhat more granular histogram, depending on the settings for `binwidth` and `dotsize`.

```
ggplot(data = nh_500cc, aes(x = Height)) +  
  geom_dotplot(dotsizes = 0.75, binwidth = 1) +  
  scale_y_continuous(NULL, breaks = NULL) + # hide y axis  
  labs(title = "Height of NHANES subjects ages 21-64",  
       x = "Height in cm.")
```

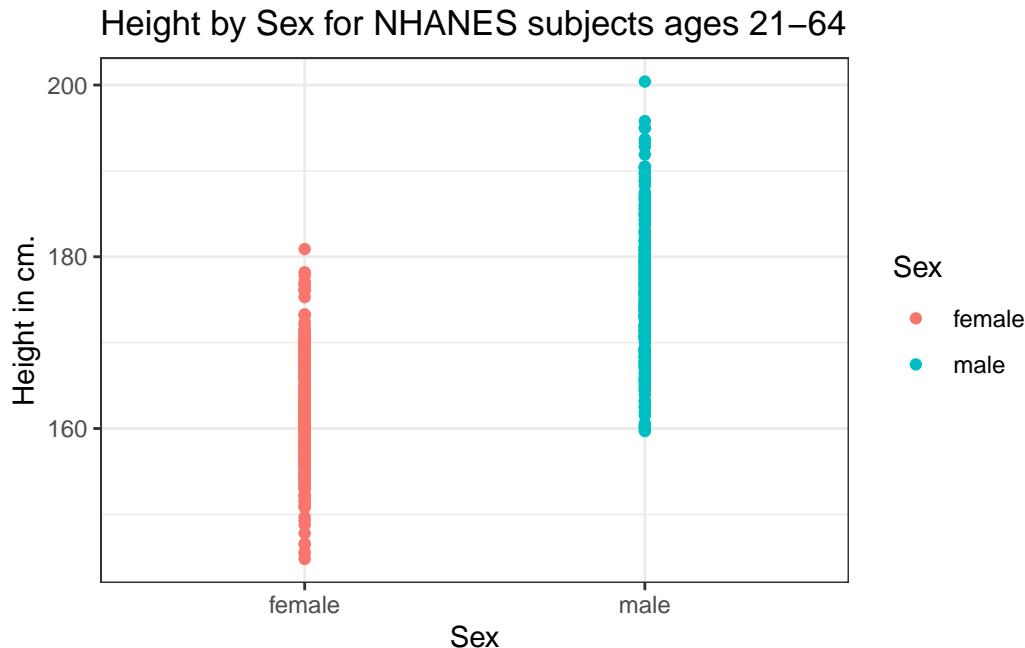
Height of NHANES subjects ages 21–64



5.4 Height and Sex

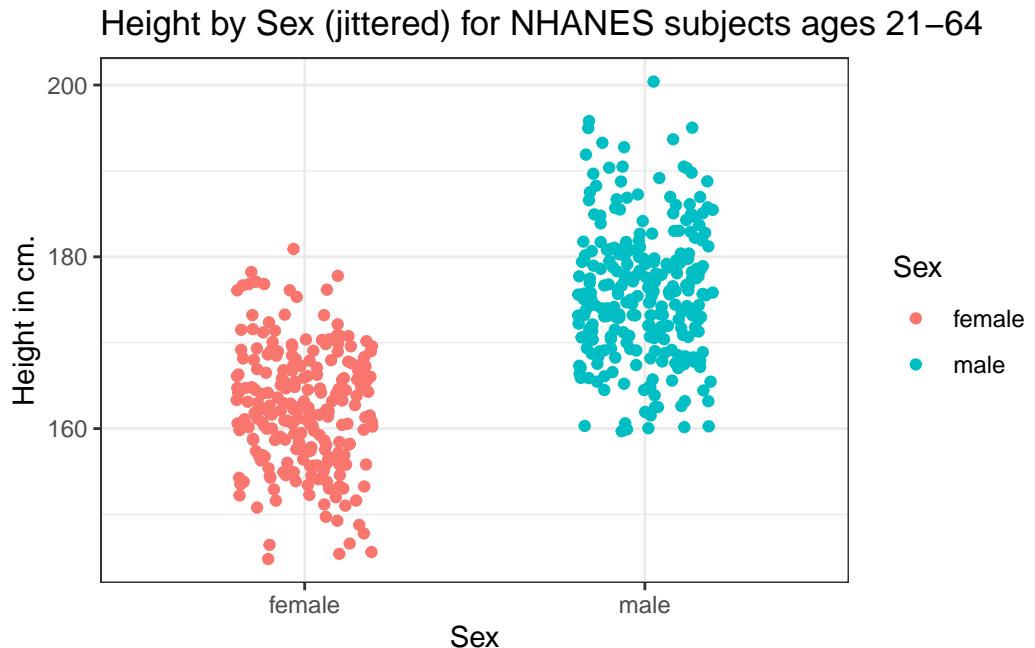
Let's look again at the impact of a respondent's sex on their height, but now within our sample of adults.

```
ggplot(data = nh_500cc,
        aes(x = Sex, y = Height, color = Sex)) +
  geom_point() +
  labs(title = "Height by Sex for NHANES subjects ages 21–64",
       y = "Height in cm.")
```



This plot isn't so useful. We can improve things a little by jittering the points horizontally, so that the overlap is reduced.

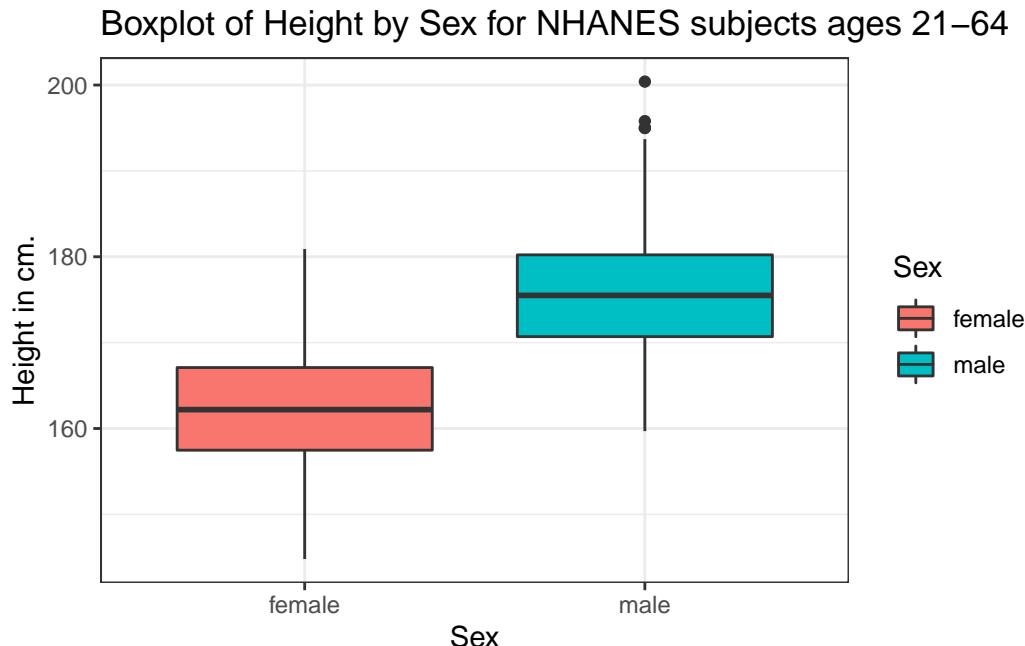
```
ggplot(data = nh_500cc, aes(x = Sex, y = Height, color = Sex)) +
  geom_jitter(width = 0.2) +
  labs(title = "Height by Sex (jittered) for NHANES subjects ages 21-64",
       y = "Height in cm.")
```



Perhaps it might be better to summarise the distribution in a different way. We might consider a boxplot of the data.

5.4.1 A Boxplot of Height by Sex

```
ggplot(data = nh_500cc, aes(x = Sex, y = Height, fill = Sex)) +
  geom_boxplot() +
  labs(title = "Boxplot of Height by Sex for NHANES subjects ages 21–64",
       y = "Height in cm.")
```



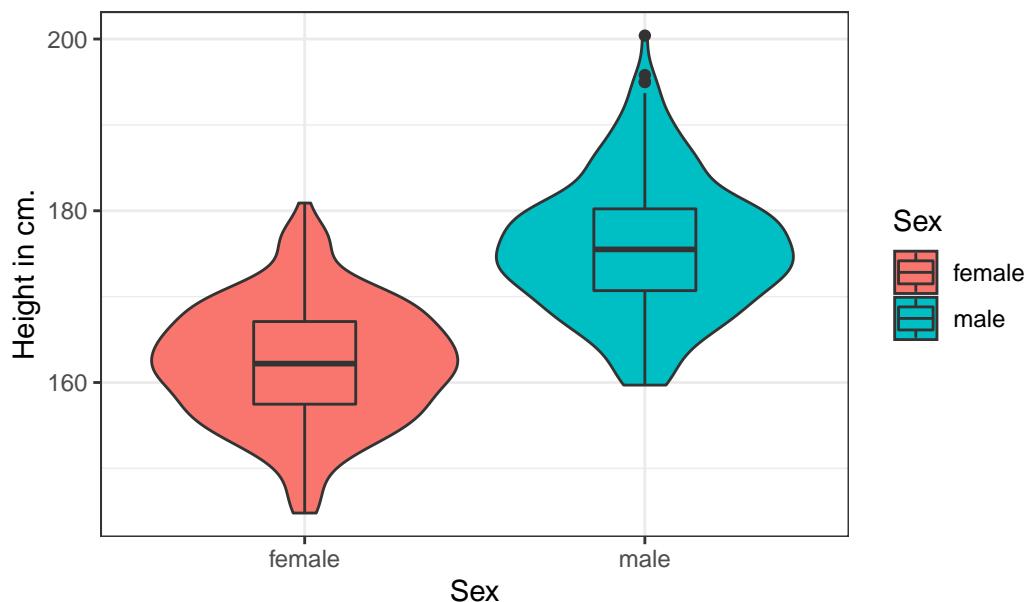
The boxplot shows some summary statistics based on percentiles. The boxes in the middle show the data values that include the middle half of the data once its been sorted. The 25th percentile (value that exceeds 1/4 of the data) is indicated by the bottom of the box, while the top of the box is located at the 75th percentile. The solid line inside the box indicates the median (also called the 50th percentile) of the Heights for that Sex.

5.4.2 Adding a violin plot

A boxplot is often supplemented with a *violin plot* to better show the shape of the distribution.

```
ggplot(data = nh_500cc, aes(x = Sex, y = Height, fill = Sex)) +
  geom_violin() +
  geom_boxplot(width = 0.3) +
  labs(title = "Boxplot of Height by Sex for NHANES subjects ages 21-64",
       y = "Height in cm.")
```

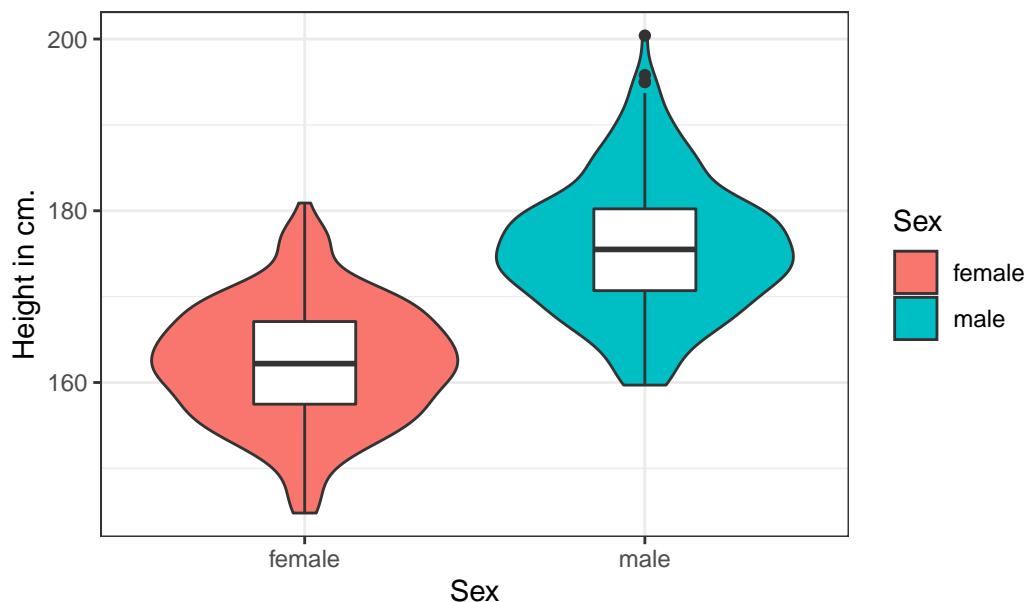
Boxplot of Height by Sex for NHANES subjects ages 21–64



This usually works better if the boxes are given a different fill than the violins, as shown in the following figure.

```
ggplot(data = nh_500cc, aes(x = Sex, y = Height)) +  
  geom_violin(aes(fill = Sex)) +  
  geom_boxplot(width = 0.3) +  
  labs(title = "Boxplot of Height by Sex for NHANES subjects ages 21–64",  
       y = "Height in cm.")
```

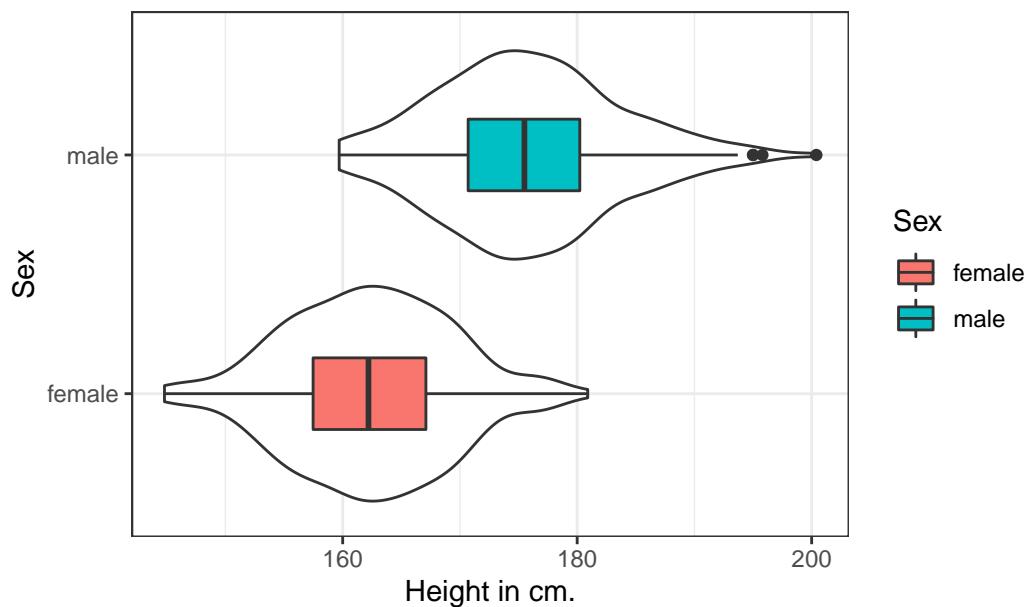
Boxplot of Height by Sex for NHANES subjects ages 21–64



We can also flip the boxplots on their side, using `coord_flip()`.

```
ggplot(data = nh_500cc, aes(x = Sex, y = Height)) +  
  geom_violin() +  
  geom_boxplot(aes(fill = Sex), width = 0.3) +  
  labs(title = "Boxplot of Height by Sex for NHANES subjects ages 21-64",  
       y = "Height in cm.") +  
  coord_flip()
```

Boxplot of Height by Sex for NHANES subjects ages 21–64

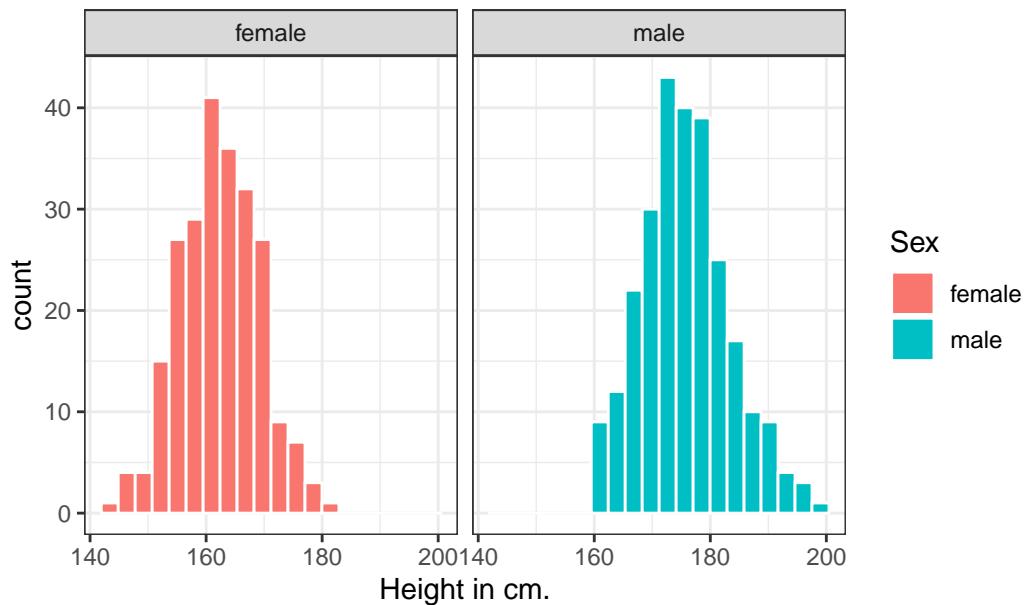


5.4.3 Histograms of Height by Sex

Or perhaps we'd like to see a pair of histograms?

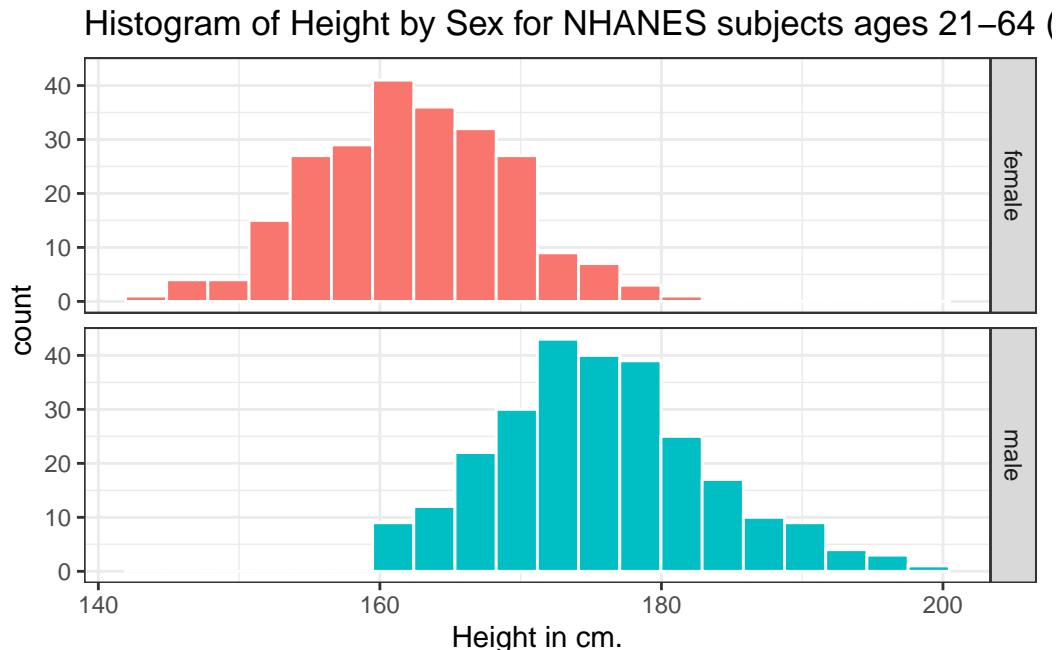
```
ggplot(data = nh_500cc, aes(x = Height, fill = Sex)) +  
  geom_histogram(color = "white", bins = 20) +  
  labs(title = "Histogram of Height by Sex for NHANES subjects ages 21-64",  
       x = "Height in cm.") +  
  facet_wrap(~ Sex)
```

Histogram of Height by Sex for NHANES subjects ages 21–64



Can we redraw these histograms so that they are a little more comparable, and to get rid of the unnecessary legend?

```
ggplot(data = nh_500cc, aes(x = Height, fill = Sex)) +  
  geom_histogram(color = "white", bins = 20) +  
  labs(title = "Histogram of Height by Sex for NHANES subjects ages 21–64 (Revised)",  
       x = "Height in cm.") +  
  guides(fill = "none") +  
  facet_grid(Sex ~ .)
```



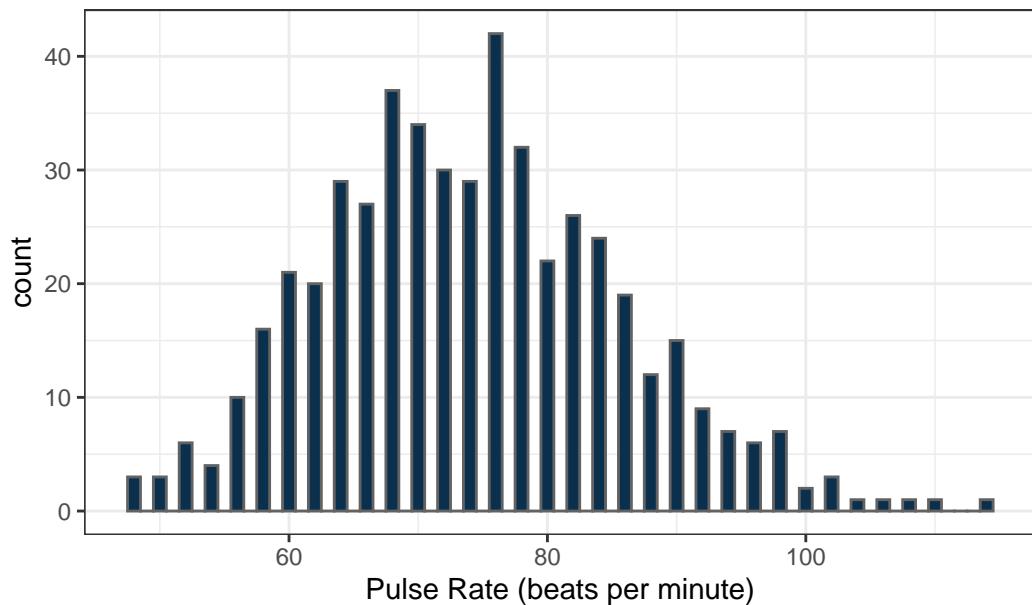
5.5 Looking at Pulse Rate

Let's look at a different outcome, the *pulse rate* for our subjects.

Here's a histogram, again with CWRU colors, for the pulse rates in our sample.

```
ggplot(data = nh_500cc, aes(x = Pulse)) +
  geom_histogram(binwidth = 1,
                 fill = cwrugray, col = cwrublue) +
  labs(title = "Histogram of Pulse Rate: NHANES subjects ages 21-64",
       x = "Pulse Rate (beats per minute)")
```

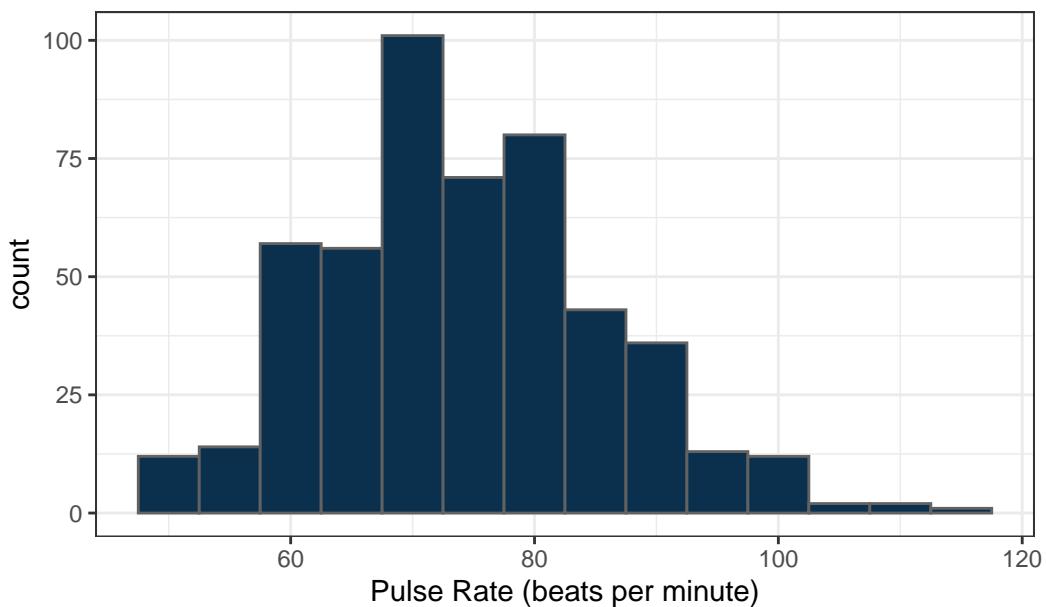
Histogram of Pulse Rate: NHANES subjects ages 21–64



Suppose we instead bin up groups of 5 beats per minute together as we plot the Pulse rates.

```
ggplot(data = nh_500cc, aes(x = Pulse)) +  
  geom_histogram(binwidth = 5,  
                 fill = cwru.blue, col = cwru.gray) +  
  labs(title = "Histogram of Pulse Rate: NHANES subjects ages 21–64",  
        x = "Pulse Rate (beats per minute)")
```

Histogram of Pulse Rate: NHANES subjects ages 21–64



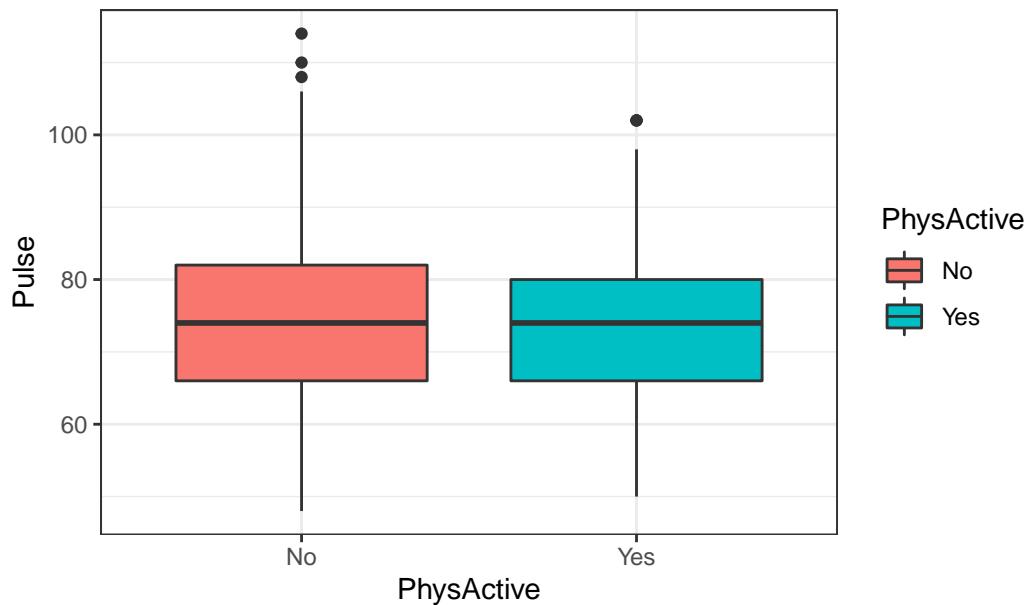
Which is the more useful representation will depend a lot on what questions you're trying to answer.

5.5.1 Pulse Rate and Physical Activity

We can also split up our data into groups based on whether the subjects are physically active. Let's try a boxplot.

```
ggplot(data = nh_500cc,  
       aes(y = Pulse, x = PhysActive, fill = PhysActive)) +  
  geom_boxplot() +  
  labs(title = "Pulse Rate by Physical Activity Status for NHANES ages 21-64")
```

Pulse Rate by Physical Activity Status for NHANES ages 21–6



As an accompanying numerical summary, we might ask how many people fall into each of these `PhysActive` categories, and what is their “average” `Pulse` rate.

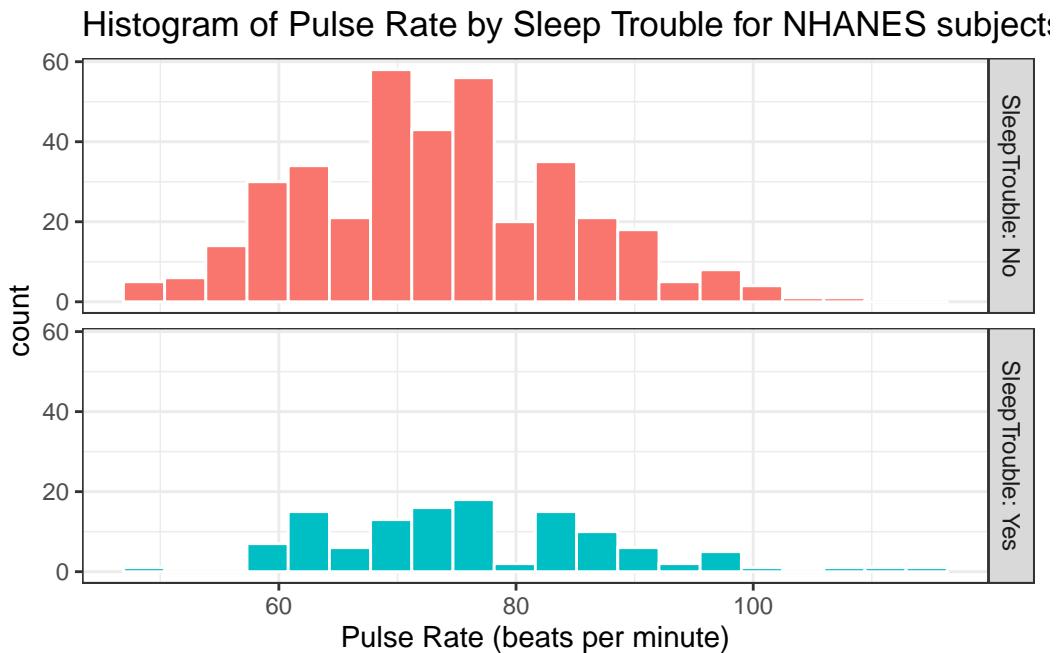
```
nh_500cc |>
  group_by(PhysActive) |>
  summarise(count = n(), mean(Pulse), median(Pulse)) |>
  kable(digits = 2)
```

PhysActive	count	mean(Pulse)	median(Pulse)
No	216	74.44	74
Yes	284	73.96	74

The `kable(digits = 2)` piece of this command tells R Markdown to generate a table with some attractive formatting, and rounding any decimals to two figures.

5.5.2 Pulse by Sleeping Trouble

```
ggplot(data = nh_500cc, aes(x = Pulse, fill = SleepTrouble)) +  
  geom_histogram(color = "white", bins = 20) +  
  labs(title = "Histogram of Pulse Rate by Sleep Trouble for NHANES subjects ages 21-64",  
        x = "Pulse Rate (beats per minute)") +  
  guides(fill = "none") +  
  facet_grid(SleepTrouble ~ ., labeller = "label_both")
```



How many people fall into each of these `SleepTrouble` categories, and what is their “average” Pulse rate?

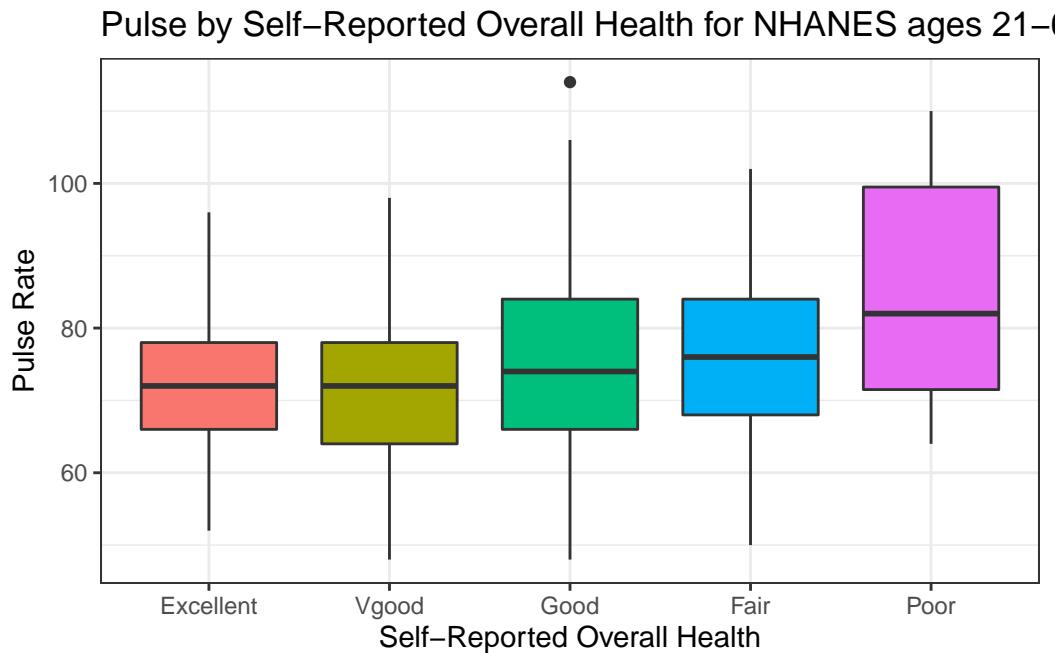
```
nh_500cc |>  
  group_by(SleepTrouble) |>  
  summarise(count = n(), mean(Pulse), median(Pulse)) |>  
  kable(digits = 2)
```

SleepTrouble	count	mean(Pulse)	median(Pulse)
No	380	73.45	73
Yes	120	76.43	76

5.5.3 Pulse and HealthGen

We can compare the distribution of Pulse rate across groups by the subject's self-reported overall health (HealthGen), as well.

```
ggplot(data = nh_500cc, aes(x = HealthGen, y = Pulse, fill = HealthGen)) +  
  geom_boxplot() +  
  labs(title = "Pulse by Self-Reported Overall Health for NHANES ages 21-64",  
       x = "Self-Reported Overall Health", y = "Pulse Rate") +  
  guides(fill = "none")
```



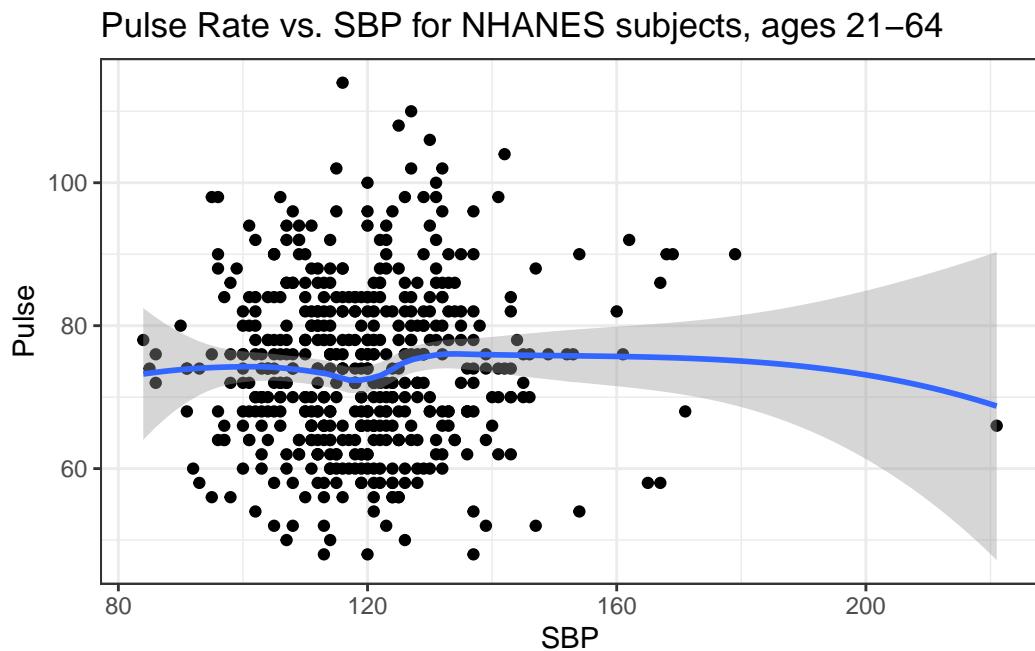
How many people fall into each of these HealthGen categories, and what is their “average” Pulse rate?

```
nh_500cc |>  
  group_by(HealthGen) |>  
  summarise(count = n(), mean(Pulse), median(Pulse)) |>  
  kable(digits = 2)
```

HealthGen	count	mean(Pulse)	median(Pulse)
Excellent	52	72.08	72
Vgood	167	71.78	72
Good	204	75.22	74
Fair	65	76.55	76
Poor	12	85.50	82

5.5.4 Pulse Rate and Systolic Blood Pressure

```
ggplot(data = nh_500cc, aes(x = SBP, y = Pulse)) +
  geom_point() +
  geom_smooth(method = "loess", formula = y ~ x) +
  labs(title = "Pulse Rate vs. SBP for NHANES subjects, ages 21-64")
```

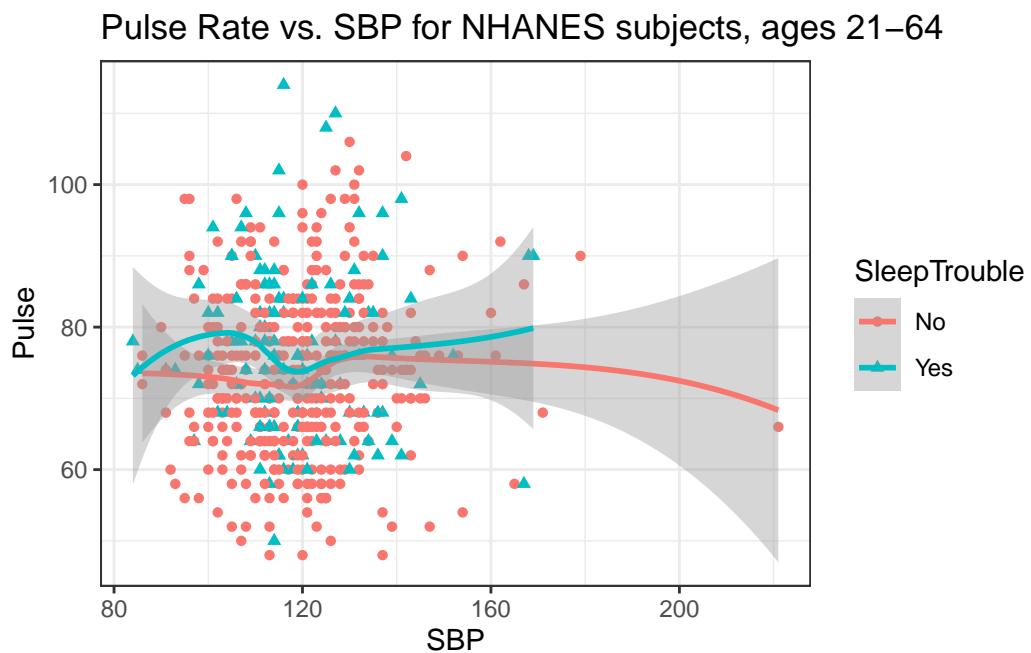


5.5.5 Sleep Trouble vs. No Sleep Trouble?

Could we see whether subjects who have described `SleepTrouble` show different SBP-pulse patterns than the subjects who haven't?

- Let's try doing this by changing the shape *and* the color of the points based on `SleepTrouble`.

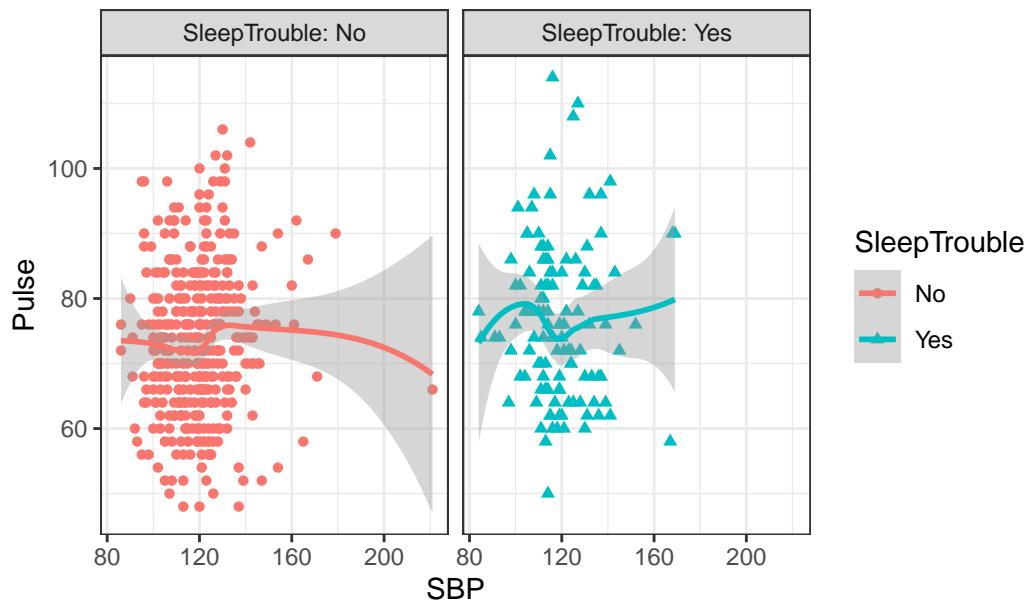
```
ggplot(data = nh_500cc,
       aes(x = SBP, y = Pulse,
           color = SleepTrouble, shape = SleepTrouble)) +
  geom_point() +
  geom_smooth(method = "loess", formula = y ~ x) +
  labs(title = "Pulse Rate vs. SBP for NHANES subjects, ages 21-64")
```



This plot might be easier to interpret if we faceted by `SleepTrouble`, as well.

```
ggplot(data = nh_500cc,
       aes(x = SBP, y = Pulse,
           color = SleepTrouble, shape = SleepTrouble)) +
  geom_point() +
  geom_smooth(method = "loess", formula = y ~ x) +
  labs(title = "Pulse Rate vs. SBP for NHANES subjects, ages 21-64") +
  facet_wrap(~ SleepTrouble, labeller = "label_both")
```

Pulse Rate vs. SBP for NHANES subjects, ages 21–64



5.6 General Health Status

Here's a Table of the General Health Status results. Again, this is a self-reported rating of each subject's health on a five point scale (Excellent, Very Good, Good, Fair, Poor.)

```
nh_500cc |>
  tabyl(HealthGen)
```

HealthGen	n	percent
Excellent	52	0.104
Vgood	167	0.334
Good	204	0.408
Fair	65	0.130
Poor	12	0.024

The HealthGen data are categorical, which means that summarizing them with averages isn't as appealing as looking at percentages, proportions and rates. The `tabyl` function comes from the `janitor` package in R.

- I don't actually like the title of `percent` here, as it's really a proportion, but that can be adjusted, and we can add a total.

```
nh_500cc |>
  tabyl(HealthGen) |>
  adorn_totals() |>
  adorn_pct_formatting()
```

HealthGen n percent

Excellent	52	10.4%
Vgood	167	33.4%
Good	204	40.8%
Fair	65	13.0%
Poor	12	2.4%
Total	500	100.0%

When working with an unordered categorical variable, like `MaritalStatus`, the same approach can work.

```
nh_500cc |>
  tabyl(MaritalStatus) |>
  adorn_totals() |>
  adorn_pct_formatting()
```

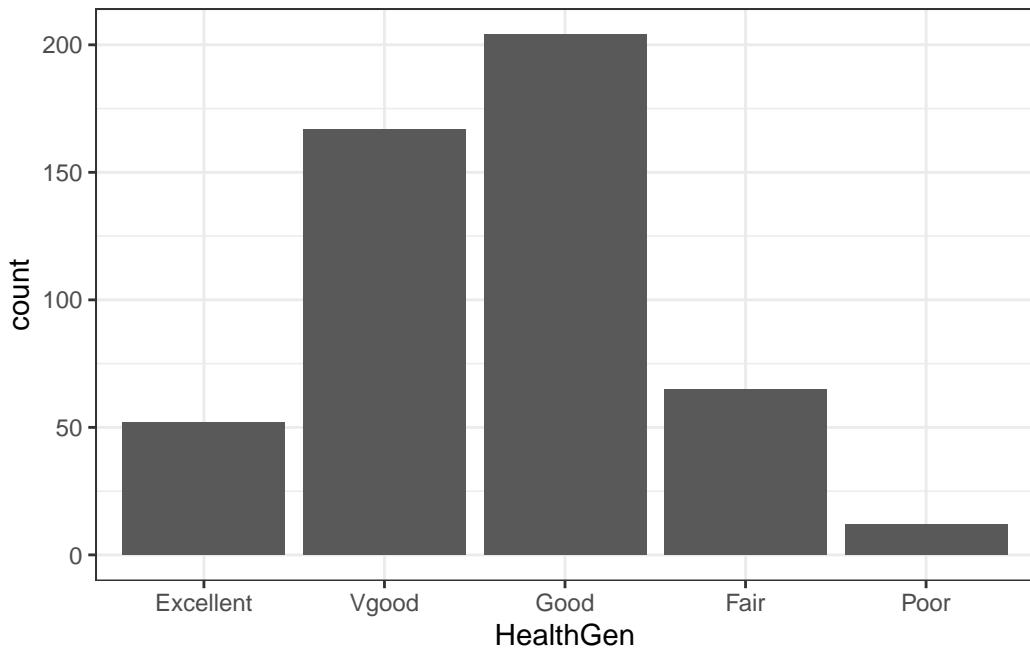
MaritalStatus n percent

Divorced	47	9.4%
LivePartner	46	9.2%
Married	256	51.2%
NeverMarried	125	25.0%
Separated	17	3.4%
Widowed	9	1.8%
Total	500	100.0%

5.6.1 Bar Chart for Categorical Data

Usually, a **bar chart** is the best choice for graphing a variable made up of categories.

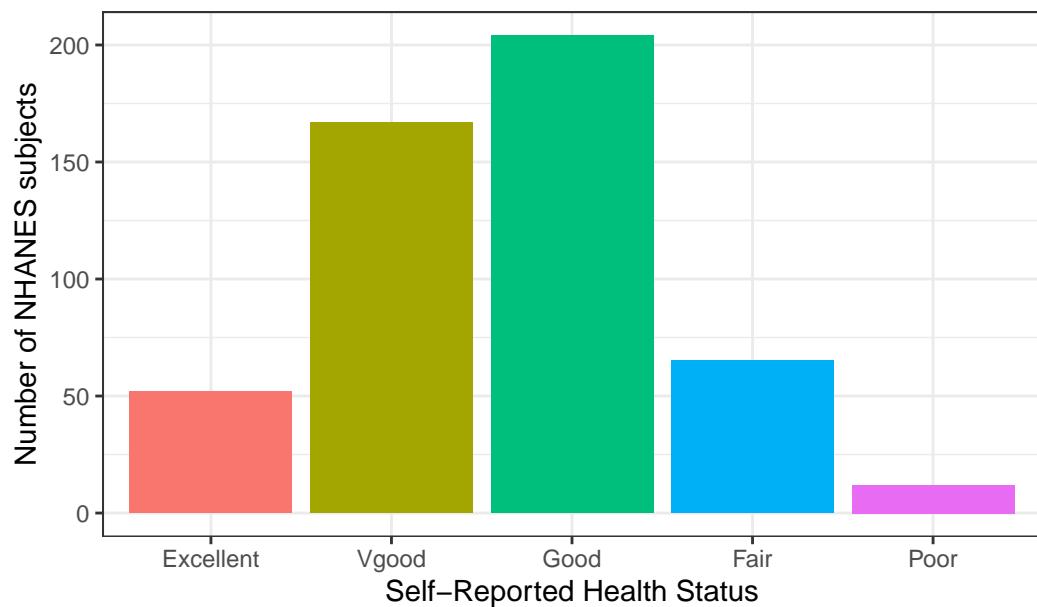
```
ggplot(data = nh_500cc, aes(x = HealthGen)) +
  geom_bar()
```



There are lots of things we can do to make this plot fancier.

```
ggplot(data = nh_500cc, aes(x = HealthGen, fill = HealthGen)) +  
  geom_bar() +  
  guides(fill = "none") +  
  labs(x = "Self-Reported Health Status",  
       y = "Number of NHANES subjects",  
       title = "Self-Reported Health Status in NHANES subjects ages 21-64")
```

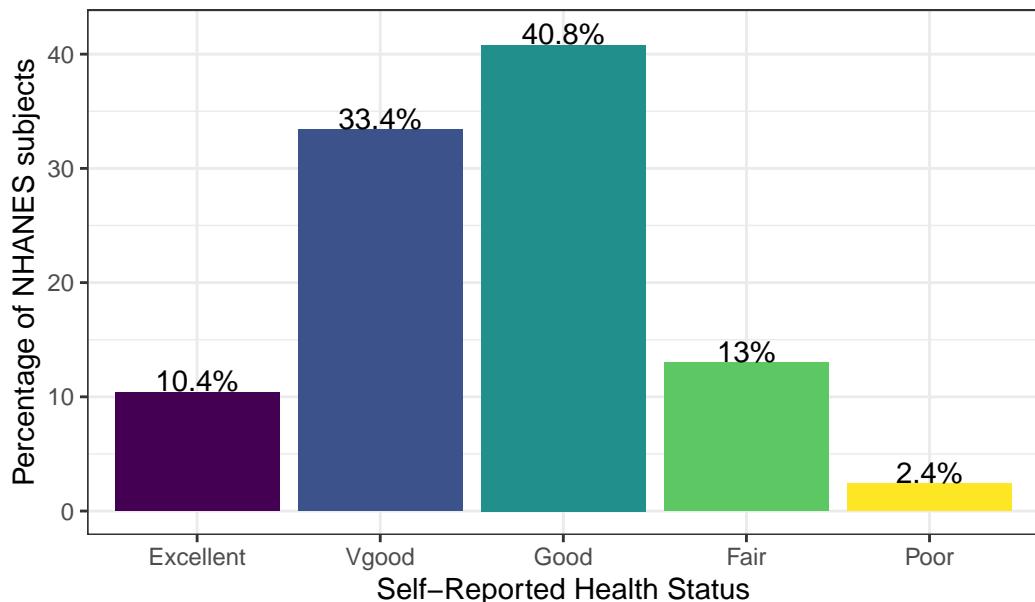
Self-Reported Health Status in NHANES subjects ages 21–64



Or, we can really go crazy...

```
nh_500cc |>
  count(HealthGen) |>
  mutate(pct = round_half_up(prop.table(n) * 100, 1)) |>
  ggplot(aes(x = HealthGen, y = pct, fill = HealthGen)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  geom_text(aes(y = pct + 1,      # nudge above top of bar
                label = paste0(pct, '%')),   # prettify
            position = position_dodge(width = .9),
            size = 4) +
  labs(x = "Self-Reported Health Status",
       y = "Percentage of NHANES subjects",
       title = "Self-Reported Health Status in NHANES subjects ages 21-64") +
  theme_bw()
```

Self-Reported Health Status in NHANES subjects ages 21–64



5.6.2 Two-Way Tables

We can create cross-classifications of two categorical variables (for example HealthGen and Smoke100), adding both row and column marginal totals, and compare subjects by Sex, as follows...

```
nh_500cc |>
  tabyl(Smoke100, HealthGen) |>
  adorn_totals(c("row", "col"))
```

Smoke100	Excellent	Vgood	Good	Fair	Poor	Total
No	44	108	105	29	5	291
Yes	8	59	99	36	7	209
Total	52	167	204	65	12	500

If we like, we can make this look a little more polished with the `knitr::kable` function...

```
nh_500cc |>
  tabyl(Smoke100, HealthGen) |>
  adorn_totals(c("row", "col")) |>
  knitr::kable()
```

	Smoke100	Excellent	Vgood	Good	Fair	Poor	Total
No	44	108	105	29	5	291	
Yes	8	59	99	36	7	209	
Total	52	167	204	65	12	500	

Or, we can get a complete cross-tabulation, including (in this case) the percentages of people within each of the two categories of `Smoke100` that fall in each `HealthGen` category (percentages within each row) like this.

```
nh_500cc |>
  tabyl(Smoke100, HealthGen) |>
  adorn_totals("row") |>
  adorn_percentages("row") |>
  adorn_pct_formatting() |>
  adorn_ns() |>
  knitr::kable()
```

Smoke100	Excellent	Vgood	Good	Fair	Poor
No	15.1% (44)	37.1% (108)	36.1% (105)	10.0% (29)	1.7% (5)
Yes	3.8% (8)	28.2% (59)	47.4% (99)	17.2% (36)	3.3% (7)
Total	10.4% (52)	33.4% (167)	40.8% (204)	13.0% (65)	2.4% (12)

And, if we wanted the column percentages, to determine which sex had the higher rate of each `HealthGen` status level, we can get that by changing the `adorn_percentages` to describe results at the column level:

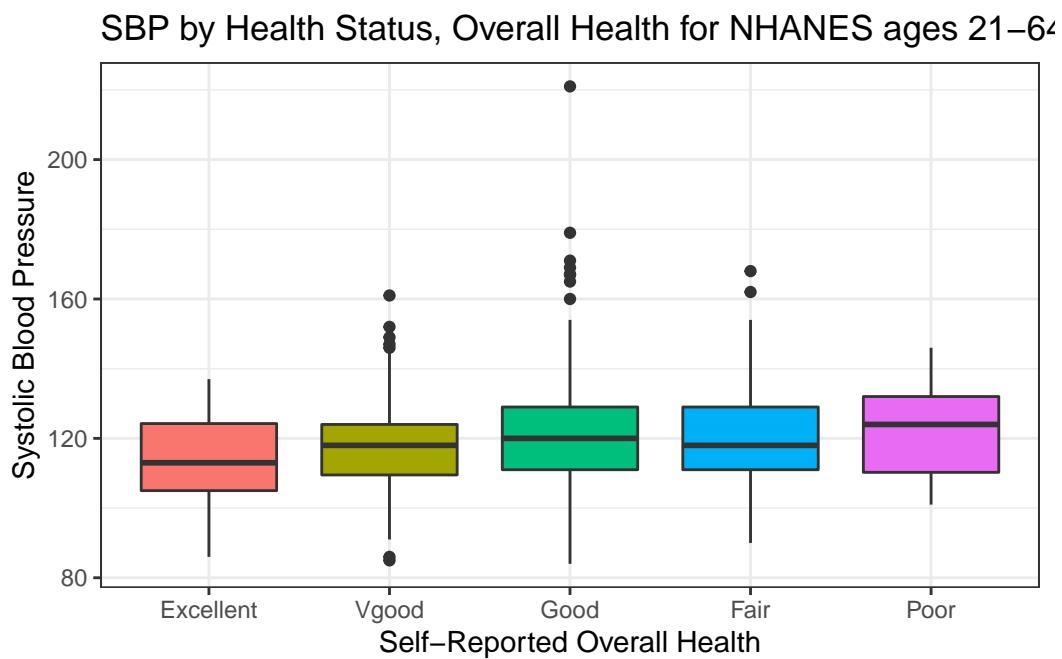
```
nh_500cc |>
  tabyl(Sex, HealthGen) |>
  adorn_totals("col") |>
  adorn_percentages("col") |>
  adorn_pct_formatting() |>
  adorn_ns() |>
  knitr::kable()
```

Sex	Excellent	Vgood	Good	Fair	Poor	Total
female	63.5% (33)	44.3% (74)	43.6% (89)	47.7% (31)	75.0% (9)	47.2% (236)
male	36.5% (19)	55.7% (93)	56.4% (115)	52.3% (34)	25.0% (3)	52.8% (264)

5.6.3 SBP by General Health Status

Let's consider now the relationship between self-reported overall health and systolic blood pressure.

```
ggplot(data = nh_500cc, aes(x = HealthGen, y = SBP,
                               fill = HealthGen)) +
  geom_boxplot() +
  labs(title = "SBP by Health Status, Overall Health for NHANES ages 21-64",
       y = "Systolic Blood Pressure",
       x = "Self-Reported Overall Health") +
  guides(fill = "none")
```



We can see that not too many people self-identify with the “Poor” health category.

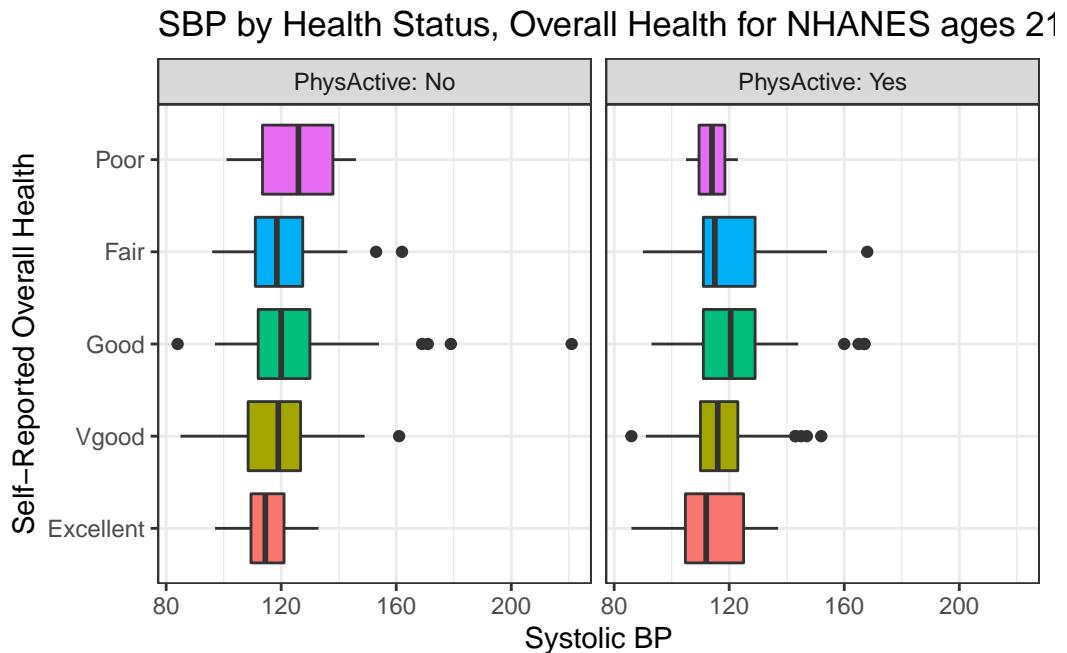
```
nh_500cc |>
  group_by(HealthGen) |>
  summarise(count = n(), mean(SBP), median(SBP)) |>
  knitr::kable()
```

HealthGen	count	mean(SBP)	median(SBP)
Excellent	52	113.9231	113
Vgood	167	117.5928	118
Good	204	121.5931	120
Fair	65	120.3846	118
Poor	12	122.8333	124

5.6.4 SBP by Physical Activity and General Health Status

We'll build a panel of boxplots to try to understand the relationships between Systolic Blood Pressure, General Health Status and Physical Activity. Note the use of `coord_flip` to rotate the graph 90 degrees, and the use of `labeler` within `facet_wrap` to include both the name of the (Physical Activity) variable and its value.

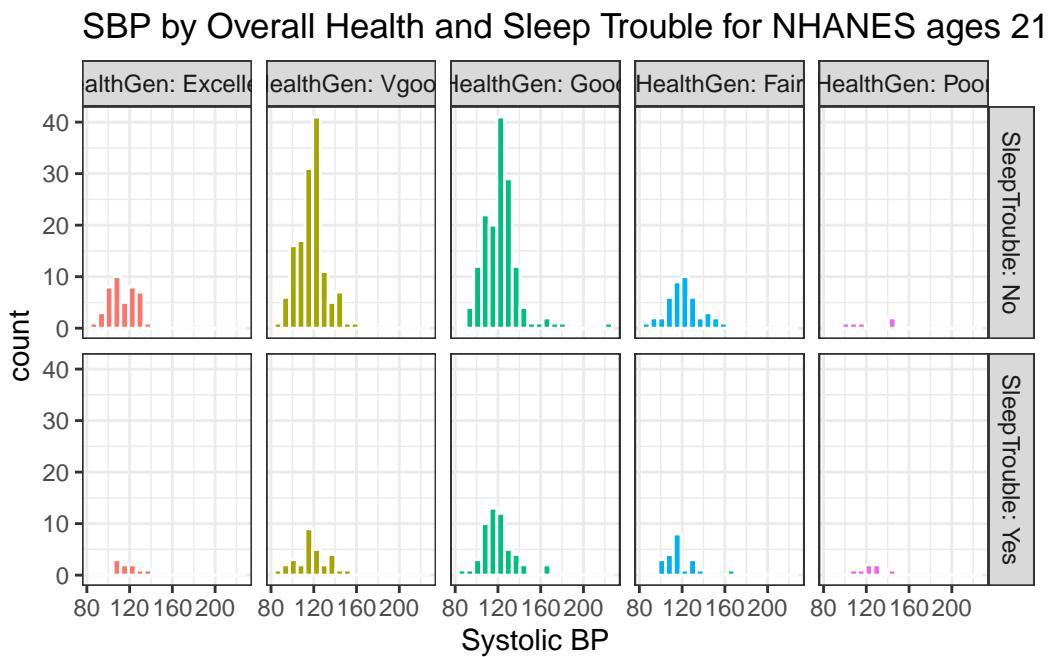
```
ggplot(data = nh_500cc, aes(x = HealthGen, y = SBP, fill = HealthGen)) +
  geom_boxplot() +
  labs(title = "SBP by Health Status, Overall Health for NHANES ages 21-64",
       y = "Systolic BP", x = "Self-Reported Overall Health") +
  guides(fill = "none") +
  facet_wrap(~ PhysActive, labeller = "label_both") +
  coord_flip()
```



5.6.5 SBP by Sleep Trouble and General Health Status

Here's a plot of faceted histograms, which might be used to address similar questions related to the relationship between Overall Health, Systolic Blood Pressure and whether someone has trouble sleeping.

```
ggplot(data = nh_500cc, aes(x = SBP, fill = HealthGen)) +  
  geom_histogram(color = "white", bins = 20) +  
  labs(title = "SBP by Overall Health and Sleep Trouble for NHANES ages 21-64",  
       x = "Systolic BP") +  
  guides(fill = "none") +  
  facet_grid(SleepTrouble ~ HealthGen, labeller = "label_both")
```



5.7 Conclusions

This is just a small piece of the toolbox for visualizations that we'll create in this class. Many additional tools are on the way, but the main idea won't change. Using the `ggplot2` package, we can accomplish several critical tasks in creating a visualization, including:

- Identifying (and labeling) the axes and titles
- Identifying a type of `geom` to use, like a point, bar or histogram

- Changing fill, color, shape, size to facilitate comparisons
- Building “small multiples” of plots with faceting

Good data visualizations make it easy to see the data, and `ggplot2`'s tools make it relatively difficult to make a really bad graph.

6 Summarizing Quantities

Most numerical summaries that might be new to you are applied most appropriately to quantitative variables. The measures that will interest us relate to:

- the **center** of our distribution,
- the **spread** of our distribution, and
- the **shape** of our distribution.

6.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

This chapter also requires that the `knitr`, `Hmisc`, `mosaic`, and `psych` packages are loaded on your machine, but are not included with `library()` above.

6.2 Working with the nh_750 data

To demonstrate key ideas in this Chapter, we will consider our sample of 750 adults ages 21-64 from NHANES 2011-12 which includes some missing values. We'll load into the `nh_750` data frame the information from the `nh_adult750.Rds` file we created in Section @ref(newNHANES).

```
nh_750 <- read_rds("data/nh_adult750.Rds")
```

6.3 The summary function for Quantitative data

R provides a small sampling of numerical summaries with the **summary** function, for instance.

```
nh_750 |>
  select(Age, BMI, SBP, DBP, Pulse) |>
  summary()
```

	Age	BMI	SBP	DBP
Min.	:21.00	Min. :16.70	Min. : 83.0	Min. : 0.00
1st Qu.	:30.00	1st Qu.:24.20	1st Qu.:108.0	1st Qu.: 66.00
Median	:40.00	Median :27.90	Median :118.0	Median : 73.00
Mean	:40.82	Mean :29.08	Mean :118.8	Mean : 72.69
3rd Qu.	:51.00	3rd Qu.:32.10	3rd Qu.:127.0	3rd Qu.: 80.00
Max.	:64.00	Max. :80.60	Max. :209.0	Max. :108.00
	NA's :5	NA's :33	NA's :33	
	Pulse			
Min.	: 40.00			
1st Qu.	: 66.00			
Median	: 72.00			
Mean	: 73.53			
3rd Qu.	: 80.00			
Max.	:124.00			
NA's	:32			

This basic summary includes a set of five **quantiles**¹, plus the sample's **mean**.

- **Min.** = the **minimum** value for each variable, so, for example, the youngest subject's Age was 21.
- **1st Qu.** = the **first quartile** (25th percentile) for each variable - for example, 25% of the subjects were Age 30 or younger.
- **Median** = the **median** (50th percentile) - half of the subjects were Age 40 or younger.
- **Mean** = the **mean**, usually what one means by an *average* - the sum of the Ages divided by 750 is 40.8,
- **3rd Qu.** = the **third quartile** (75th percentile) - 25% of the subjects were Age 51 or older.
- **Max.** = the **maximum** value for each variable, so the oldest subject was Age 64.

The summary also specifies the number of missing values for each variable. Here, we are missing 5 of the BMI values, for example.

¹The quantiles (sometimes referred to as percentiles) can also be summarized with a boxplot.

6.4 Measuring the Center of a Distribution

6.4.1 The Mean and The Median

The **mean** and **median** are the most commonly used measures of the center of a distribution for a quantitative variable. The median is the more generally useful value, as it is relevant even if the data have a shape that is not symmetric. We might also collect the **sum** of the observations, and the **count** of the number of observations, usually symbolized with n .

For variables without missing values, like `Age`, this is pretty straightforward.

```
nh_750 |>
  summarise(n = n(), Mean = mean(Age), Median = median(Age), Sum = sum(Age))

# A tibble: 1 x 4
  n  Mean Median   Sum
<int> <dbl>  <dbl> <int>
1    750   40.8    40 30616
```

And again, the Mean is just the Sum (30616), divided by the number of non-missing values of Age (750), or 40.8213333.

The Median is the middle value when the data are sorted in order. When we have an odd number of values, this is sufficient. When we have an even number, as in this case, we take the mean of the two middle values. We could sort and list all 500 Ages, if we wanted to do so.

```
nh_750 |> select(Age) |>
  arrange(Age)

# A tibble: 750 x 1
  Age
  <int>
1 21
2 21
3 21
4 21
5 21
6 21
7 21
8 21
```

```

9    21
10   21
# ... with 740 more rows
# i Use `print(n = ...)` to see more rows

```

But this data set figures we don't want to output more than 10 observations to a table like this.

If we really want to see all of the data, we can use `View(nh_750)` to get a spreadsheet-style presentation, or use the `sort` command...

```
sort(nh_750$Age)
```

```

[1] 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 21 22 22 22
[26] 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 22 23 23 23 23 23 23
[51] 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 23 24 24 24 24 24 24 24 24
[76] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 25 25 25 25 25 25 25 25 25 25
[101] 25 25 25 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 27 27 27 27 27 27 27 27
[126] 27 27 27 27 27 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28 28
[151] 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 30 30 30 30
[176] 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 31 31 31 31 31 31
[201] 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 32 32 32 32 32 32 32 32 32 32
[226] 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 33 33 33 33 33 33 33 33
[251] 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 33 34 34 34 34 34 34 34 34 34
[276] 34 34 34 34 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 36 36 36 36 36 36 36 36 36
[301] 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 37 37 37 37 37 37 37 37 37
[326] 37 37 37 37 37 37 37 37 37 37 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38
[351] 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 40 40 40 40
[376] 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 41 41 41 41 41 41 41 41 41 41 41 41 41 41
[401] 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 43 43 43 43 43 43 43 43
[426] 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 44 44 44 44 44 44 44 44 44 44 44 44 44 44
[451] 44 44 44 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 46 46 46 46
[476] 46 46 46 46 46 46 46 46 46 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47
[501] 47 48 48 48 48 48 48 48 48 48 48 48 48 48 49 49 49 49 49 49 49 49 49 49 49 49 49 49 49 49
[526] 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50
[551] 50 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 51 52 52 52
[576] 52 52 52 52 52 52 52 52 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 54 54
[601] 54 54 54 54 54 54 54 54 54 54 54 54 54 54 54 54 55 55 55 55 55 55 55 55 55 55 55 55 55 56
[626] 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 57 57 57 57 57 57 57
[651] 57 57 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 59 59 59 59 59
[676] 59 59 59 59 59 59 59 59 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 61 61 61 61
[701] 61 61 61 61 61 61 61 61 62 62 62 62 62 62 62 62 62 62 62 62 62 62 62 62 62 63 63

```

```
[726] 63 63 63 63 63 63 63 63 63 63 63 63 64 64 64 64 64 64 64 64 64 64 64 64 64 64
```

Again, to find the median, we would take the mean of the middle two observations in this sorted data set. That would be the 250th and 251st largest Ages.

```
sort(nh_750$Age) [250:251]
```

```
[1] 33 33
```

6.4.2 Dealing with Missingness

When calculating a mean, you may be tempted to try something like this...

```
nh_750 |>
  summarise(mean(Pulse), median(Pulse))

# A tibble: 1 x 2
`mean(Pulse)` `median(Pulse)`
<dbl>          <int>
1           NA            NA
```

This fails because we have some missing values in the Pulse data. We can address this by either omitting the data with missing values before we run the `summarise()` function, or tell the mean and median summary functions to remove missing values².

```
nh_750 |>
  filter(complete.cases(Pulse)) |>
  summarise(count = n(), mean(Pulse), median(Pulse))

# A tibble: 1 x 3
count `mean(Pulse)` `median(Pulse)`
<int>      <dbl>        <dbl>
1    718       73.5        72
```

Or, we could tell the summary functions themselves to remove NA values.

²We could also use `!is.na` in place of `complete.cases` to accomplish the same thing.

```

nh_750 |>
  summarise(mean(Pulse, na.rm=TRUE), median(Pulse, na.rm=TRUE))

# A tibble: 1 x 2
`mean(Pulse, na.rm = TRUE)` `median(Pulse, na.rm = TRUE)`
<dbl>                  <dbl>
1                      73.5                 72

```

In Chapter @ref(miss), we will discuss various assumptions we can make about missing data, and the importance of **imputation** when dealing with it in modeling or making inferences. For now, we will limit our descriptive summaries to observed values, in what are called complete case or available case analyses.

6.4.3 The Mode of a Quantitative Variable

One other less common measure of the center of a quantitative variable's distribution is its most frequently observed value, referred to as the **mode**. This measure is only appropriate for discrete variables, be they quantitative or categorical. To find the mode, we usually tabulate the data, and then sort by the counts of the numbers of observations.

```

nh_750 |>
  group_by(Age) |>
  summarise(count = n()) |>
  arrange(desc(count))

# A tibble: 44 x 2
  Age   count
  <int> <int>
1 32     28
2 36     26
3 50     26
4 30     24
5 33     24
6 24     23
7 21     22
8 22     22
9 23     22
10 28    20
# ... with 34 more rows
# i Use `print(n = ...)` to see more rows

```

The mode is just the most common Age observed in the data.

Note the use of three different “verbs” in our function there - for more explanation of this strategy, visit Wickham and Grolemund (2022). The `group_by` function here is very useful. It converts the `nh_750` data frame into a new grouped tibble where operations are performed on the groups. Here, this means that it groups the data by Age before counting observations, and then sorting the groups (the Ages) by their frequencies.

As an alternative, the `modeest` package’s `mfv` function calculates the sample mode (or most frequent value)³.

6.5 Measuring the Spread of a Distribution

Statistics is all about variation, so spread or dispersion is an important fundamental concept in statistics. Measures of spread like the inter-quartile range and range (maximum - minimum) can help us understand and compare data sets. If the values in the data are close to the center, the spread will be small. If many of the values in the data are scattered far away from the center, the spread will be large.

6.5.1 The Range and the Interquartile Range (IQR)

The `range` of a quantitative variable is sometimes interpreted as the difference between the maximum and the minimum, even though R presents the actual minimum and maximum values when you ask for a range...

```
nh_750 |>
  select(Age) |>
  range()
```

```
[1] 21 64
```

And, for a variable with missing values, we can use...

```
nh_750 |>
  select(BMI) |>
  filter(complete.cases(BMI)) |>
  range()
```

³See the documentation for the `modeest` package’s `mfv` function to look at other definitions of the mode.

```
[1] 16.7 80.6
```

A more interesting and useful statistic is the **inter-quartile range**, or IQR, which is the range of the middle half of the distribution, calculated by subtracting the 25th percentile value from the 75th percentile value.

```
nh_750 |>
  summarise(IQR(Age), quantile(Age, 0.25), quantile(Age, 0.75))
```

```
# A tibble: 1 x 3
`IQR(Age)` `quantile(Age, 0.25)` `quantile(Age, 0.75)`
<dbl>          <dbl>          <dbl>
1       21            30            51
```

We can calculate the range and IQR nicely from the summary information on quantiles, of course:

```
nh_750 |>
  select(Age, BMI, SBP, DBP, Pulse) |>
  summary()
```

	Age	BMI	SBP	DBP
Min.	:21.00	Min. :16.70	Min. : 83.0	Min. : 0.00
1st Qu.	:30.00	1st Qu.:24.20	1st Qu.:108.0	1st Qu.: 66.00
Median	:40.00	Median :27.90	Median :118.0	Median : 73.00
Mean	:40.82	Mean :29.08	Mean :118.8	Mean : 72.69
3rd Qu.	:51.00	3rd Qu.:32.10	3rd Qu.:127.0	3rd Qu.: 80.00
Max.	:64.00	Max. :80.60	Max. :209.0	Max. :108.00
		NA's :5	NA's :33	NA's :33

	Pulse
Min.	: 40.00
1st Qu.	: 66.00
Median	: 72.00
Mean	: 73.53
3rd Qu.	: 80.00
Max.	:124.00
NA's	:32

6.5.2 The Variance and the Standard Deviation

The IQR is always a reasonable summary of spread, just as the median is always a reasonable summary of the center of a distribution. Yet, most people are inclined to summarize a batch of data using two numbers: the **mean** and the **standard deviation**. This is really only a sensible thing to do if you are willing to assume the data follow a Normal distribution: a bell-shaped, symmetric distribution without substantial outliers.

But **most data do not (even approximately) follow a Normal distribution**. Summarizing by the median and quartiles (25th and 75th percentiles) is much more robust, explaining R's emphasis on them.

6.5.3 Obtaining the Variance and Standard Deviation in R

Here are the variances of the quantitative variables in the `nh_750` data. Note the need to include `na.rm = TRUE` to deal with the missing values in some variables.

```
nh_750 |>
  select(Age, BMI, SBP, DBP, Pulse) |>
  summarise_all(var, na.rm = TRUE)

# A tibble: 1 x 5
  Age    BMI   SBP   DBP Pulse
  <dbl> <dbl> <dbl> <dbl> <dbl>
1 157.  52.4  229.  128.  136.
```

And here are the standard deviations of those same variables.

```
nh_750 |>
  select(Age, BMI, SBP, DBP, Pulse) |>
  summarise_all(sd, na.rm = TRUE)

# A tibble: 1 x 5
  Age    BMI   SBP   DBP Pulse
  <dbl> <dbl> <dbl> <dbl> <dbl>
1 12.5  7.24  15.1  11.3  11.6
```

6.5.4 Defining the Variance and Standard Deviation

Bock, Velleman, and De Veaux (2004) have lots of useful thoughts here, which are lightly edited here.

In thinking about spread, we might consider how far each data value is from the mean. Such a difference is called a *deviation*. We could just average the deviations, but the positive and negative differences always cancel out, leaving an average deviation of zero, so that's not helpful. Instead, we *square* each deviation to obtain non-negative values, and to emphasize larger differences. When we add up these squared deviations and find their mean (almost), this yields the **variance**.

$$\text{Variance} = s^2 = \frac{\sum(y - \bar{y})^2}{n - 1}$$

Why almost? It would be the mean of the squared deviations only if we divided the sum by n , but instead we divide by $n - 1$ because doing so produces an estimate of the true (population) variance that is *unbiased*⁴. If you're looking for a more intuitive explanation, [this Stack Exchange link](#) awaits your attention.

- To return to the original units of measurement, we take the square root of s^2 , and instead work with s , the **standard deviation**, also abbreviated SD.

$$\text{Standard Deviation} = s = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}$$

6.5.5 Interpreting the SD when the data are Normally distributed

For a set of measurements that follow a Normal distribution, the interval:

- Mean \pm Standard Deviation contains approximately 68% of the measurements;
- Mean $\pm 2(\text{Standard Deviation})$ contains approximately 95% of the measurements;
- Mean $\pm 3(\text{Standard Deviation})$ contains approximately all (99.7%) of the measurements.

We often refer to the population or process mean of a distribution with μ and the standard deviation with σ , leading to the Figure below.

But if the data are not from an approximately Normal distribution, then this Empirical Rule is less helpful.

⁴When we divide by $n-1$ as we calculate the sample variance, the average of the sample variances for all possible samples is equal to the population variance. If we instead divided by n , the average sample variance across all possible samples would be a little smaller than the population variance.

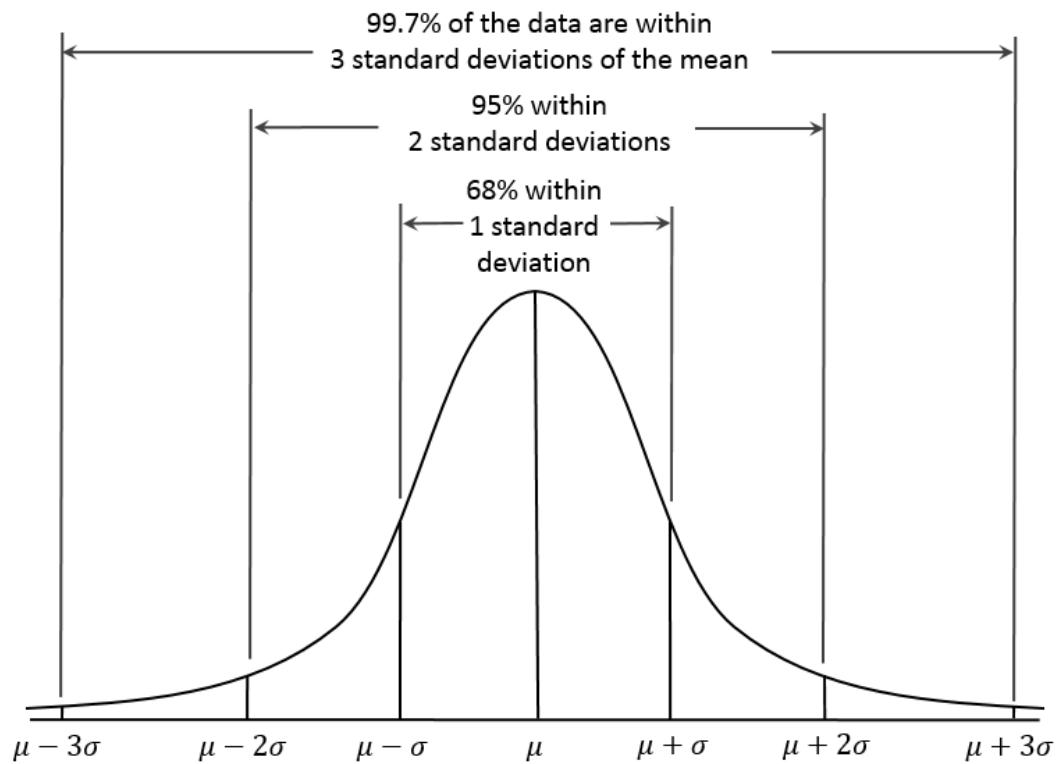


Figure 6.1: The Normal Distribution and the Empirical Rule

6.5.6 Chebyshev's Inequality: One Interpretation of the Standard Deviation

Chebyshev's Inequality tells us that for any distribution, regardless of its relationship to a Normal distribution, no more than $1/k^2$ of the distribution's values can lie more than k standard deviations from the mean. This implies, for instance, that for **any** distribution, at least 75% of the values must lie within two standard deviations of the mean, and at least 89% must lie within three standard deviations of the mean.

Again, most data sets do not follow a Normal distribution. We'll return to this notion soon. But first, let's try to draw some pictures that let us get a better understanding of the distribution of our data.

6.6 Measuring the Shape of a Distribution

When considering the shape of a distribution, one is often interested in three key points.

- The number of modes in the distribution, which I always assess through plotting the data.
- The **skewness**, or symmetry that is present, which I typically assess by looking at a plot of the distribution of the data, but if required to, will summarize with a non-parametric measure of **skewness**.
- The **kurtosis**, or heavy-tailedness (outlier-proneness) that is present, usually in comparison to a Normal distribution. Again, this is something I nearly inevitably assess graphically, but there are measures.

A Normal distribution has a single mode, is symmetric and, naturally, is neither heavy-tailed nor light-tailed as compared to a Normal distribution (we call this mesokurtic).

6.6.1 Multimodal vs. Unimodal distributions

A unimodal distribution, on some level, is straightforward. It is a distribution with a single mode, or “peak” in the distribution. Such a distribution may be skewed or symmetric, light-tailed or heavy-tailed. We usually describe as multimodal distributions like the two on the right below, which have multiple local maxima, even though they have just a single global maximum peak.

Truly multimodal distributions are usually described that way in terms of shape. For unimodal distributions, skewness and kurtosis become useful ideas.

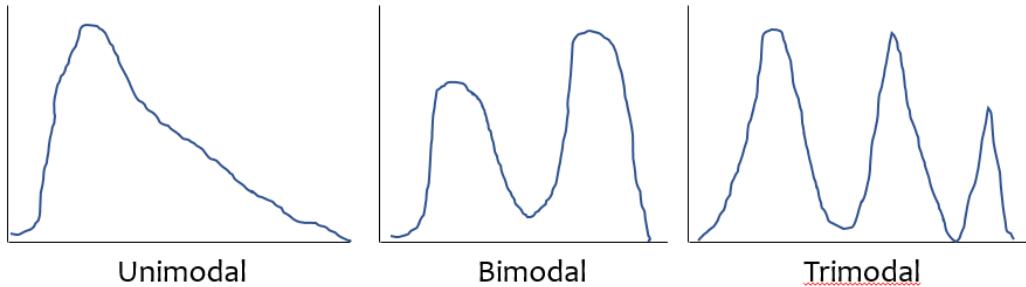


Figure 6.2: Unimodal and Multimodal Sketches

6.6.2 Skew

Whether or not a distribution is approximately symmetric is an important consideration in describing its shape. Graphical assessments are always most useful in this setting, particularly for unimodal data. My favorite measure of skew, or skewness if the data have a single mode, is:

$$skew_1 = \frac{\text{mean} - \text{median}}{\text{standard deviation}}$$

- Symmetric distributions generally show values of $skew_1$ near zero. If the distribution is actually symmetric, the mean should be equal to the median.
- Distributions with $skew_1$ values above 0.2 in absolute value generally indicate meaningful skew.
- Positive skew (mean > median if the data are unimodal) is also referred to as *right skew*.
- Negative skew (mean < median if the data are unimodal) is referred to as *left skew*.

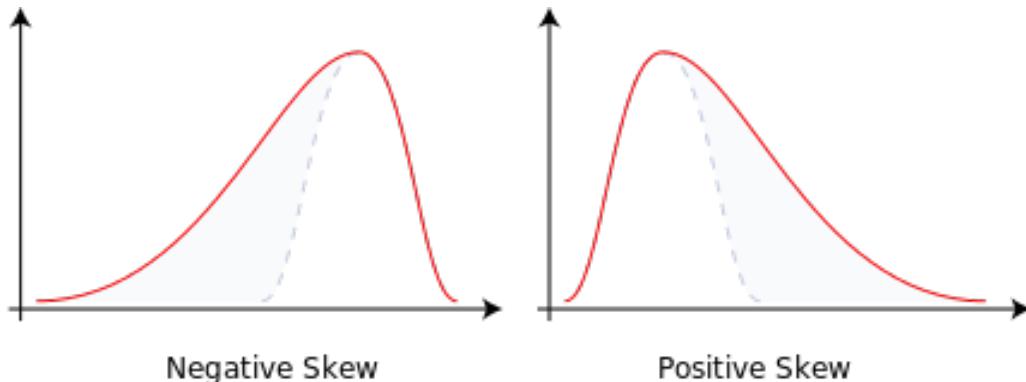


Figure 6.3: Negative (Left) Skew and Positive (Right) Skew

6.6.3 Kurtosis

When we have a unimodal distribution that is symmetric, we will often be interested in the behavior of the tails of the distribution, as compared to a Normal distribution with the same mean and standard deviation. High values of kurtosis measures (and there are several) indicate data which has extreme outliers, or is heavy-tailed.

- A mesokurtic distribution has similar tail behavior to what we would expect from a Normal distribution.
- A leptokurtic distribution is a thinner, more slender distribution, with heavier tails than we'd expect from a Normal distribution. One example is the t distribution.
- A platykurtic distribution is a broader, flatter distribution, with thinner tails than we'd expect from a Normal distribution. One example is a uniform distribution.

```
set.seed(431)
sims_kurt <- tibble(meso = rnorm(n = 300, mean = 0, sd = 1),
                     lepto = rt(n = 300, df = 4),
                     platy = runif(n = 300, min = -2, max = 2))

p1 <- ggplot(sims_kurt, aes(x = meso)) +
  geom_histogram(aes(y = stat(density)),
                 bins = 25, fill = "royalblue", col = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean(sims_kurt$meso),
                            sd = sd(sims_kurt$meso)),
                col = "red") +
  labs(title = "Normal (mesokurtic)")

p1a <- ggplot(sims_kurt, aes(x = meso, y = "")) +
  geom_violin() +
  geom_boxplot(fill = "royalblue", outlier.color = "royalblue", width = 0.3) +
  labs(y = "", x = "Normal (mesokurtic)")

p2 <- ggplot(sims_kurt, aes(x = lepto)) +
  geom_histogram(aes(y = stat(density)),
                 bins = 25, fill = "tomato", col = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean(sims_kurt$lepto),
                            sd = sd(sims_kurt$lepto)),
                col = "royalblue") +
  labs(title = "t (leptokurtic)")
```

```

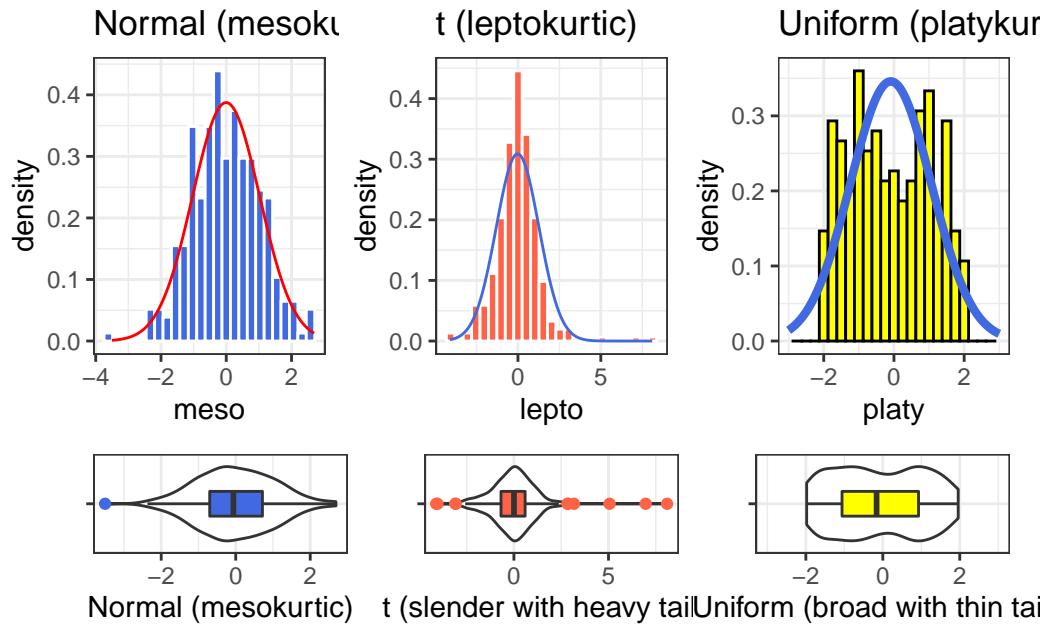
p2a <- ggplot(sims_kurt, aes(x = lepto, y = "")) +
  geom_violin() +
  geom_boxplot(fill = "tomato", outlier.color = "tomato", width = 0.3) +
  labs(y = "", x = "t (slender with heavy tails)")

p3 <- ggplot(sims_kurt, aes(x = platy)) +
  geom_histogram(aes(y = stat(density)),
                 bins = 25, fill = "yellow", col = "black") +
  stat_function(fun = dnorm,
                args = list(mean = mean(sims_kurt$platy),
                            sd = sd(sims_kurt$platy)),
                col = "royalblue", lwd = 1.5) +
  xlim(-3, 3) +
  labs(title = "Uniform (platykurtic)")

p3a <- ggplot(sims_kurt, aes(x = platy, y = "")) +
  geom_violin() +
  geom_boxplot(fill = "yellow", width = 0.3) +
  xlim(-3, 3) +
  labs(y = "", x = "Uniform (broad with thin tails)")

(p1 + p2 + p3) / (p1a + p2a + p3a) +
  plot_layout(heights = c(3, 1))

```



Graphical tools are in most cases the best way to identify issues related to kurtosis.

6.7 Numerical Summaries for Quantitative Variables

6.7.1 favstats in the mosaic package

The `favstats` function adds the standard deviation, and counts of overall and missing observations to our usual `summary` for a continuous variable. Let's look at systolic blood pressure, because we haven't yet.

```
mosaic::favstats(~ SBP, data = nh_750)

Registered S3 method overwritten by 'mosaic':
  method                  from
  fortify.SpatialPolygonsDataFrame ggplot2

  min   Q1 median   Q3 max      mean       sd    n missing
  83 108    118 127 209 118.7908 15.14329 717      33
```

We could, of course, duplicate these results with several `summarise()` pieces...

```

nh_750 |>
  filter(complete.cases(SBP)) |>
  summarise(min = min(SBP), Q1 = quantile(SBP, 0.25),
            median = median(SBP), Q3 = quantile(SBP, 0.75),
            max = max(SBP), mean = mean(SBP),
            sd = sd(SBP), n = n(), miss = sum(is.na(SBP)))

# A tibble: 1 x 9
  min     Q1 median     Q3   max   mean     sd     n miss
  <int> <dbl> <int> <dbl> <int> <dbl> <dbl> <int> <int>
1    83    108    118    127    209   119.   15.1    717     0

```

The somewhat unusual structure of `favstats` (complete with an easy to forget `~`) is actually helpful. It allows you to look at some interesting grouping approaches, like this:

```
mosaic::favstats(SBP ~ Education, data = nh_750)
```

	Education	min	Q1	median	Q3	max	mean	sd	n	missing
1	8th Grade	96	110.25	119.5	129.75	167	122.4565	16.34993	46	4
2	9 - 11th Grade	85	107.75	116.0	127.00	191	118.8026	15.79453	76	0
3	High School	84	111.50	120.5	129.00	209	121.0882	16.52853	136	7
4	Some College	85	108.00	117.0	126.00	186	118.6293	14.32736	232	9
5	College Grad	83	107.00	117.0	125.00	171	116.8326	14.41202	227	13

Of course, we could accomplish the same comparison with `dplyr` commands, too, but the `favstats` approach has much to offer.

```

nh_750 |>
  filter(complete.cases(SBP, Education)) |>
  group_by(Education) |>
  summarise(min = min(SBP), Q1 = quantile(SBP, 0.25),
            median = median(SBP), Q3 = quantile(SBP, 0.75),
            max = max(SBP), mean = mean(SBP),
            sd = sd(SBP), n = n(), miss = sum(is.na(SBP)))

# A tibble: 5 x 10
  Education      min     Q1 median     Q3   max   mean     sd     n miss
  <fct>       <int> <dbl> <dbl> <dbl> <int> <dbl> <dbl> <int> <int>
1 8th Grade      96    110.    120.    130.    167   122.   16.3    46     0

```

2 9 - 11th Grade	85	108.	116	127	191	119.	15.8	76	0
3 High School	84	112.	120.	129	209	121.	16.5	136	0
4 Some College	85	108	117	126	186	119.	14.3	232	0
5 College Grad	83	107	117	125	171	117.	14.4	227	0

6.7.2 describe in the psych package

The `psych` package has a more detailed list of numerical summaries for quantitative variables that lets us look at a group of observations at once.

```
psych::describe(nh_750 |> select(Age, BMI, SBP, DBP, Pulse))
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
Age	1	750	40.82	12.54	40.0	40.53	14.83	21.0	64.0	43.0	0.16
BMI	2	745	29.08	7.24	27.9	28.31	5.93	16.7	80.6	63.9	1.72
SBP	3	717	118.79	15.14	118.0	117.88	13.34	83.0	209.0	126.0	0.96
DBP	4	717	72.69	11.34	73.0	72.65	10.38	0.0	108.0	108.0	-0.28
Pulse	5	718	73.53	11.65	72.0	73.11	11.86	40.0	124.0	84.0	0.48
		kurtosis	se								
Age		-1.15	0.46								
BMI		6.16	0.27								
SBP		3.10	0.57								
DBP		2.59	0.42								
Pulse		0.73	0.43								

The additional statistics presented here are:

- `trimmed` = a trimmed mean (by default in this function, this removes the top and bottom 10% from the data, then computes the mean of the remaining values - the middle 80% of the full data set.)
- `mad` = the median absolute deviation (from the median), which can be used in a manner similar to the standard deviation or IQR to measure spread.
 - If the data are Y_1, Y_2, \dots, Y_n , then the `mad` is defined as $\text{median}(|Y_i - \text{median}(Y_i)|)$.
 - To find the `mad` for a set of numbers, find the median, subtract the median from each value and find the absolute value of that difference, and then find the median of those absolute differences.
 - For non-normal data with a skewed shape but tails well approximated by the Normal, the `mad` is likely to be a better (more robust) estimate of the spread than is the standard deviation.

- a measure of `skew`, which refers to how much asymmetry is present in the shape of the distribution. The measure is not the same as the *nonparametric skew* measure that we will usually prefer. The [Wikipedia page on skewness][<https://en.wikipedia.org/wiki/Skewness>] is very detailed.
- a measure of excess `kurtosis`, which refers to how outlier-prone, or heavy-tailed the shape of the distribution is, as compared to a Normal distribution.
- `se` = the standard error of the sample mean, equal to the sample `sd` divided by the square root of the sample size.

6.7.3 The Hmisc package's version of `describe`

```
Hmisc::describe(nh_750 |>
  select(Age, BMI, SBP, DBP, Pulse))

select(nh_750, Age, BMI, SBP, DBP, Pulse)

5 Variables      750 Observations
-----
Age
  n   missing  distinct    Info     Mean     Gmd     .05     .10
  750        0       44  0.999   40.82   14.46    22     24
  .25       .50       .75    .90     .95
  30        40       51     59     62

lowest : 21 22 23 24 25, highest: 60 61 62 63 64
-----
BMI
  n   missing  distinct    Info     Mean     Gmd     .05     .10
  745        5       250     1   29.08   7.538   20.22  21.30
  .25       .50       .75    .90     .95
  24.20    27.90    32.10  37.60   41.28

lowest : 16.7 17.6 17.8 17.9 18.0, highest: 59.1 62.8 63.3 69.0 80.6
-----
SBP
  n   missing  distinct    Info     Mean     Gmd     .05     .10
  717        33       81  0.999  118.8   16.36   98.0   102.0
  .25       .50       .75    .90     .95
  108.0    118.0    127.0  137.0  144.2

lowest : 83 84 85 86 89, highest: 171 179 186 191 209
```

```

-----  

DBP  

      n   missing  distinct    Info     Mean     Gmd     .05     .10
    717       33       66  0.999  72.69  12.43    55     59
    .25       .50       .75    .90    .95
    66       73       80     86    91  

lowest :  0  25  41  42  44, highest: 104 105 106 107 108
-----  

Pulse  

      n   missing  distinct    Info     Mean     Gmd     .05     .10
    718       32       37  0.997  73.53  12.95    56     60
    .25       .50       .75    .90    .95
    66       72       80     88    94  

lowest :  40  44  46  48  50, highest: 108 112 114 118 124
-----
```

The `Hmisc` package's version of `describe` for a distribution of data presents three new ideas, in addition to a more comprehensive list of quartiles (the 5th, 10th, 25th, 50th, 75th, 90th and 95th are shown) and the lowest and highest few observations. These are:

- `distinct` - the number of different values observed in the data.
- `Info` - a measure of how “continuous” the variable is, related to how many “ties” there are in the data, with `Info` taking a higher value (closer to its maximum of one) if the data are more continuous.
- `Gmd` - the Gini mean difference - a robust measure of spread that is calculated as the mean absolute difference between any pairs of observations. Larger values of `Gmd` indicate more spread-out distributions. (Gini is pronounced as either “Genie” or “Ginny”.)

6.7.4 Other options

The package [summarytools](#) has a function called `dfSummary` which I like and Dominic Comtois has also published [Recommendations for Using summarytools with R Markdown](#). Note that this isn't really for Word documents.

[DataExplorer](#) can be used for more automated exploratory data analyses (and some people also like [skimr](#)) and [visdat](#), as well.

The `df_stats` function available when the `mosaic` package is loaded allows you to run `favstats` for multiple outcome variables simultaneously.

7 Summarizing Categories

7.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(janitor)
library(gt)
library(tidyverse)

theme_set(theme_bw())
```

7.2 Using the nh_adult750 data again

To demonstrate key ideas in this Chapter, we will again consider our sample of 750 adults ages 21-64 from NHANES 2011-12 which includes some missing values. We'll load into the `nh_750` data frame the information from the `nh_adult750.Rds` file we created in Section @ref(newNHANES).

```
nh_750 <- read_rds("data/nh_adult750.Rds")
```

Summarizing categorical variables numerically is mostly about building tables, and calculating percentages or proportions. We'll save our discussion of modeling categorical data for later. Recall that in the `nh_750` data set we built in Section @ref(newNHANES) we had the following categorical variables. The number of levels indicates the number of possible categories for each categorical variable.

Variable	Description	Levels	Type
Sex	sex of subject	2	binary
Race	subject's race	6	nominal
Education	subject's educational level	5	ordinal
PhysActive	Participates in sports?	2	binary
Smoke100	Smoked 100+ cigarettes?	2	binary

Variable	Description	Levels	Type
SleepTrouble	Trouble sleeping?	2	binary
HealthGen	Self-report health	5	ordinal

7.3 The summary function for Categorical data

When R recognizes a variable as categorical, it stores it as a *factor*. Such variables get special treatment from the `summary` function, in particular a table of available values (so long as there aren't too many.)

```
nh_750 |>
  select(Sex, Race, Education, PhysActive, Smoke100,
         SleepTrouble, HealthGen, MaritalStatus) |>
  summary()
```

Sex	Race	Education	PhysActive	Smoke100
female:388	Asian : 70	8th Grade : 50	No :326	No :453
male :362	Black :128	9 - 11th Grade: 76	Yes:424	Yes:297
	Hispanic: 63	High School :143		
	Mexican : 80	Some College :241		
	White :393	College Grad :240		
	Other : 16			
SleepTrouble	HealthGen	MaritalStatus		
No :555	Excellent: 84	Divorced : 78		
Yes:195	Vgood :197	LivePartner : 70		
	Good :252	Married :388		
	Fair :104	NeverMarried:179		
	Poor : 14	Separated : 19		
	NA's : 99	Widowed : 16		

7.4 Tables to describe One Categorical Variable

Suppose we build a table (using the `tabyl` function from the `janitor` package) to describe the `HealthGen` distribution.

```
nh_750 |>
  tabyl(HealthGen) |>
  adorn_pct_formatting()
```

HealthGen	n	percent	valid_percent
Excellent	84	11.2%	12.9%
Vgood	197	26.3%	30.3%
Good	252	33.6%	38.7%
Fair	104	13.9%	16.0%
Poor	14	1.9%	2.2%
<NA>	99	13.2%	-

Note how the missing (<NA>) values are not included in the `valid_percent` calculation, but are in the `percent` calculation. Note also the use of percentage formatting.

What if we want to add a total count, sometimes called the *marginal* total?

```
nh_750 |>
  tabyl(HealthGen) |>
  adorn_totals() |>
  adorn_pct_formatting()
```

HealthGen	n	percent	valid_percent
Excellent	84	11.2%	12.9%
Vgood	197	26.3%	30.3%
Good	252	33.6%	38.7%
Fair	104	13.9%	16.0%
Poor	14	1.9%	2.2%
<NA>	99	13.2%	-
Total	750	100.0%	100.0%

What about marital status, which has no missing data in our sample?

```
nh_750 |>
  tabyl(MaritalStatus) |>
  adorn_totals() |>
  adorn_pct_formatting()
```

MaritalStatus	n	percent
Divorced	78	10.4%
LivePartner	70	9.3%
Married	388	51.7%
NeverMarried	179	23.9%
Separated	19	2.5%
Widowed	16	2.1%
Total	750	100.0%

7.5 Constructing Tables Well

The prolific Howard Wainer is responsible for many interesting books on visualization and related issues, including Wainer (2005) and Wainer (2013). These rules come from Chapter 10 of Wainer (1997).

1. Order the rows and columns in a way that makes sense.
2. Round, a lot!
3. ALL is different and important

7.5.1 Alabama First!

Which of these Tables is more useful to you?

2013 Percent of Students in grades 9-12 who are obese

State	% Obese	95% CI	Sample Size
Alabama	17.1	(14.6 - 19.9)	1,499
Alaska	12.4	(10.5-14.6)	1,167
Arizona	10.7	(8.3-13.6)	1,520
Arkansas	17.8	(15.7-20.1)	1,470
Connecticut	12.3	(10.2-14.7)	2,270
Delaware	14.2	(12.9-15.6)	2,475
Florida	11.6	(10.5-12.8)	5,491
...			
Wisconsin	11.6	(9.7-13.9)	2,771
Wyoming	10.7	(9.4-12.2)	2,910

or ...

State	% Obese	95% CI	Sample Size
Kentucky	18.0	(15.7 - 20.6)	1,537
Arkansas	17.8	(15.7 - 20.1)	1,470
Alabama	17.1	(14.6 - 19.9)	1,499
Tennessee	16.9	(15.1 - 18.8)	1,831
Texas	15.7	(13.9 - 17.6)	3,039
...			
Massachusetts	10.2	(8.5 - 12.1)	2,547
Idaho	9.6	(8.2 - 11.1)	1,841
Montana	9.4	(8.4 - 10.5)	4,679
New Jersey	8.7	(6.8 - 11.2)	1,644

State	% Obese	95% CI	Sample Size
Utah	6.4	(4.8 - 8.5)	2,136

It is a rare event when Alabama first is the best choice.

7.5.2 ALL is different and important

Summaries of rows and columns provide a measure of what is typical or usual. Sometimes a sum is helpful, at other times, consider presenting a median or other summary. The ALL category, as Wainer (1997) suggests, should be both visually different from the individual entries and set spatially apart.

On the whole, it's *far* easier to fall into a good graph in R (at least if you have some ggplot2 skills) than to produce a good table.

7.6 The Mode of a Categorical Variable

A common measure applied to a categorical variable is to identify the mode, the most frequently observed value. To find the mode for variables with lots of categories (so that the `summary` may not be sufficient), we usually tabulate the data, and then sort by the counts of the numbers of observations, as we did with discrete quantitative variables.

```

nh_750 |>
  group_by(HealthGen) |>
  summarise(count = n()) |>
  arrange(desc(count))

# A tibble: 6 x 2
  HealthGen count
  <fct>     <int>
1 Good       252
2 Vgood      197
3 Fair        104
4 <NA>        99
5 Excellent   84
6 Poor        14

```

7.7 describe in the Hmisc package

```
Hmisc::describe(nh_750 |>
  select(Sex, Race, Education, PhysActive,
         Smoke100, SleepTrouble,
         HealthGen, MaritalStatus))

select(nh_750, Sex, Race, Education, PhysActive, Smoke100, SleepTrouble, HealthGen, MaritalS

  8 Variables      750 Observations
-----
Sex
  n  missing distinct
  750      0        2

  Value      female    male
  Frequency     388     362
  Proportion   0.517   0.483
-----
Race
  n  missing distinct
  750      0        6

  lowest : Asian      Black      Hispanic Mexican  White
  highest: Black      Hispanic Mexican  White      Other
  Value      Asian     Black Hispanic Mexican  White     Other
  Frequency    70      128      63       80      393      16
  Proportion  0.093   0.171   0.084   0.107   0.524   0.021
-----
Education
  n  missing distinct
  750      0        5

  lowest : 8th Grade      9 - 11th Grade High School      Some College  College Grad
  highest: 8th Grade      9 - 11th Grade High School      Some College  College Grad
  Value      8th Grade 9 - 11th Grade High School      Some College
  Frequency    50          76          143          241
  Proportion  0.067      0.101      0.191      0.321
```

Value	College Grad
Frequency	240
Proportion	0.320

PhysActive

n	missing	distinct
750	0	2

Value	No	Yes
Frequency	326	424
Proportion	0.435	0.565

Smoke100

n	missing	distinct
750	0	2

Value	No	Yes
Frequency	453	297
Proportion	0.604	0.396

SleepTrouble

n	missing	distinct
750	0	2

Value	No	Yes
Frequency	555	195
Proportion	0.74	0.26

HealthGen

n	missing	distinct
651	99	5

lowest : Excellent	Vgood	Good	Fair	Poor
highest: Excellent	Vgood	Good	Fair	Poor

Value	Excellent	Vgood	Good	Fair	Poor
Frequency	84	197	252	104	14
Proportion	0.129	0.303	0.387	0.160	0.022

MaritalStatus

n	missing	distinct
750	0	6

```

lowest : Divorced      LivePartner   Married       NeverMarried Separated
highest: LivePartner   Married       NeverMarried Separated     Widowed

Value          Divorced  LivePartner        Married  NeverMarried    Separated
Frequency      78        70                388      179            19
Proportion    0.104    0.093            0.517    0.239          0.025

Value          Widowed
Frequency      16
Proportion    0.021
-----
```

7.8 Cross-Tabulations of Two Variables

It is very common for us to want to describe the association of one categorical variable with another. For instance, is there a relationship between Education and SleepTrouble in these data?

```

nh_750 |>
  tabyl(Education, SleepTrouble) |>
  adorn_totals(where = c("row", "col"))
```

Education	No	Yes	Total
8th Grade	40	10	50
9 - 11th Grade	52	24	76
High School	102	41	143
Some College	173	68	241
College Grad	188	52	240
Total	555	195	750

Note the use of `adorn_totals` to get the marginal counts, and how we specify that we want both the row and column totals. We can add a title for the columns with...

```

nh_750 |>
  tabyl(Education, SleepTrouble) |>
  adorn_totals(where = c("row", "col")) |>
  adorn_title(placement = "combined")
```

Education/SleepTrouble	No	Yes	Total
------------------------	----	-----	-------

	8th Grade	40	10	50
9 - 11th Grade	52	24	76	
	High School	102	41	143
	Some College	173	68	241
	College Grad	188	52	240
	Total	555	195	750

Often, we'll want to show percentages in a cross-tabulation like this. To get row percentages so that we can directly see the probability of `SleepTrouble = Yes` for each level of `Education`, we can use:

```
nh_750 |>
  tabyl(Education, SleepTrouble) |>
  adorn_totals(where = "row") |>
  adorn_percentages(denominator = "row") |>
  adorn_pct_formatting() |>
  adorn_title(placement = "combined")
```

Education/SleepTrouble	No	Yes
8th Grade	80.0%	20.0%
9 - 11th Grade	68.4%	31.6%
High School	71.3%	28.7%
Some College	71.8%	28.2%
College Grad	78.3%	21.7%
Total	74.0%	26.0%

If we want to compare the distribution of `Education` between the two levels of `SleepTrouble` with column percentages, we can use the following...

```
nh_750 |>
  tabyl(Education, SleepTrouble) |>
  adorn_totals(where = "col") |>
  adorn_percentages(denominator = "col") |>
  adorn_pct_formatting() |>
  adorn_title(placement = "combined")
```

Education/SleepTrouble	No	Yes	Total
8th Grade	7.2%	5.1%	6.7%
9 - 11th Grade	9.4%	12.3%	10.1%
High School	18.4%	21.0%	19.1%

```
Some College 31.2% 34.9% 32.1%
College Grad 33.9% 26.7% 32.0%
```

If we want overall percentages in the cells of the table, so that the total across all combinations of Education and SleepTrouble is 100%, we can use:

```
nh_750 |>
  tabyl(Education, SleepTrouble) |>
  adorn_totals(where = c("row", "col")) |>
  adorn_percentages(denominator = "all") |>
  adorn_pct_formatting() |>
  adorn_title(placement = "combined")
```

Education/SleepTrouble	No	Yes	Total
8th Grade	5.3%	1.3%	6.7%
9 - 11th Grade	6.9%	3.2%	10.1%
High School	13.6%	5.5%	19.1%
Some College	23.1%	9.1%	32.1%
College Grad	25.1%	6.9%	32.0%
Total	74.0%	26.0%	100.0%

Another common approach is to include both counts and percentages in a cross-tabulation. Let's look at the breakdown of HealthGen by MaritalStatus.

```
nh_750 |>
  tabyl(MaritalStatus, HealthGen) |>
  adorn_totals(where = c("row")) |>
  adorn_percentages(denominator = "row") |>
  adorn_pct_formatting() |>
  adorn_ns(position = "front") |>
  adorn_title(placement = "combined") |>
  knitr::kable()
```

MaritalStatus/HealthGen	Excellent	Vgood	Good	Fair	Poor	NA_
Divorced	7 (9.0%)	19 (24.4%)	29 (37.2%)	11 (14.1%)	3 (3.8%)	9 (11.5%)
LivePartner	4 (5.7%)	19 (27.1%)	25 (35.7%)	18 (25.7%)	0 (0.0%)	4 (5.7%)
Married	46 (11.9%)	101 (26.0%)	130 (33.5%)	41 (10.6%)	6 (1.5%)	64 (16.5%)

MaritalStatus/HealthGen	Excellent	Vgood	Good	Fair	Poor	NA_
NeverMarried	25 (14.0%)	52 (29.1%)	56 (31.3%)	24 (13.4%)	3 (1.7%)	19 (10.6%)
Separated	2 (10.5%)	3 (15.8%)	4 (21.1%)	8 (42.1%)	0 (0.0%)	2 (10.5%)
Widowed	0 (0.0%)	3 (18.8%)	8 (50.0%)	2 (12.5%)	2 (12.5%)	1 (6.2%)
Total	84 (11.2%)	197 (26.3%)	252 (33.6%)	104 (13.9%)	14 (1.9%)	99 (13.2%)

What if we wanted to ignore the missing `HealthGen` values? Most often, I filter down to the complete observations.

```
nh_750 |>
  filter(complete.cases(MaritalStatus, HealthGen)) |>
  tabyl(MaritalStatus, HealthGen) |>
  adorn_totals(where = c("row")) |>
  adorn_percentages(denominator = "row") |>
  adorn_pct_formatting() |>
  adorn_ns(position = "front") |>
  adorn_title(placement = "combined")
```

MaritalStatus/HealthGen	Excellent	Vgood	Good	Fair
Divorced	7 (10.1%)	19 (27.5%)	29 (42.0%)	11 (15.9%)
LivePartner	4 (6.1%)	19 (28.8%)	25 (37.9%)	18 (27.3%)
Married	46 (14.2%)	101 (31.2%)	130 (40.1%)	41 (12.7%)
NeverMarried	25 (15.6%)	52 (32.5%)	56 (35.0%)	24 (15.0%)
Separated	2 (11.8%)	3 (17.6%)	4 (23.5%)	8 (47.1%)
Widowed	0 (0.0%)	3 (20.0%)	8 (53.3%)	2 (13.3%)
Total	84 (12.9%)	197 (30.3%)	252 (38.7%)	104 (16.0%)
Poor				
3 (4.3%)				
0 (0.0%)				
6 (1.9%)				
3 (1.9%)				
0 (0.0%)				
2 (13.3%)				
14 (2.2%)				

For more on working with `tabyls`, see the vignette in the `janitor` package. There you'll find a complete list of all of the `adorn` functions, for example.

Here's another approach, to look at the cross-classification of Race and HealthGen:

```
xtabs(~ Race + HealthGen, data = nh_750)
```

Race	HealthGen				
	Excellent	Vgood	Good	Fair	Poor
Asian	10	17	24	6	1
Black	15	28	40	24	4
Hispanic	4	9	24	13	2
Mexican	6	12	25	21	2
White	48	128	131	37	5
Other	1	3	8	3	0

7.9 Cross-Classifying Three Categorical Variables

Suppose we are interested in `Smoke100` and its relationship to `PhysActive` and `SleepTrouble`.

```
nh_750 |>
  tabyl(Smoke100, PhysActive, SleepTrouble) |>
  adorn_title(placement = "top")
```

\$No

Smoke100	PhysActive	
	No	Yes
No	137	219
Yes	93	106

\$Yes

Smoke100	PhysActive	
	No	Yes
No	41	56
Yes	55	43

The result here is a `tabyl` of `Smoke100` (rows) by `PhysActive` (columns), split into a list by `SleepTrouble`.

There are several alternative approaches for doing this, although I expect us to stick with `tabyl` for our work in 431. These alternatives include the use of the `xtabs` function:

```
xtabs(~ Smoke100 + PhysActive + SleepTrouble, data = nh_750)
```

```
, , SleepTrouble = No
```

```
    PhysActive
Smoke100  No Yes
  No   137 219
  Yes   93 106
```

```
, , SleepTrouble = Yes
```

```
    PhysActive
Smoke100  No Yes
  No    41  56
  Yes   55  43
```

We can also build a **flat** version of this table, as follows:

```
ftable(Smoke100 ~ PhysActive + SleepTrouble, data = nh_750)
```

		Smoke100	
		No	Yes
PhysActive	SleepTrouble		
	No	137	93
	Yes	41	55
SleepTrouble	No	219	106
	Yes	56	43

And we can do this with **dplyr** functions and the **table()** function, as well, for example...

```
nh_750 |>
  select(Smoke100, PhysActive, SleepTrouble) |>
  table()
```

```
, , SleepTrouble = No
```

```
    PhysActive
Smoke100  No Yes
  No   137 219
  Yes   93 106
```

```
, , SleepTrouble = Yes
```

```
    PhysActive
```

Smoke		100	No	Yes
No	41	56		
Yes	55	43		

7.10 Gaining Control over Tables in R: the `gt` package

With the `gt` package, anyone can make wonderful-looking tables using the R programming language. The `gt` package allows you to start with a tibble or data frame, and use it to make very detailed tables that look professional, and includes tools that enable you to include titles and subtitles, all sorts of labels, as well as footnotes and source notes.

Here's a fairly simple example of a cross-tabulation of part of the `nh_750` data built using a few tools from the `gt` package.

```
temp_tbl <- nh_750 |> filter(complete.cases(PhysActive, HealthGen)) |>
  tabyl(PhysActive, HealthGen) |>
  tibble()

gt(temp_tbl) |>
  tab_header(title = md("##Cross-Tabulation from nh_750"),
             subtitle = md("Physical Activity vs. Overall Health"))
```

Cross-Tabulation from nh_750						
Physical Activity vs. Overall Health						
PhysActive	Excellent	Vgood	Good	Fair	Poor	
No	24	66	126	59	10	
Yes	60	131	126	45	4	

The `gt` package and its usage is described in detail at <https://gt.rstudio.com/>.

8 Missing Data and Single Imputation

Almost all serious statistical analyses have to deal with missing data. Data values that are missing are indicated in R, and to R, by the symbol NA.

8.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(janitor)
library(naniar)
library(simputation)
library(tidyverse)

theme_set(theme_bw())
```

We'll focus on tools from the `naniar` and `simputation` packages in this chapter. This chapter also requires that the `mosaic` package is loaded on your machine, but that package is not included with `library()` above.

8.2 A Simulated Example with 15 subjects

In the following tiny data set called `sbp_example`, we have four variables for a set of 15 subjects. In addition to a subject id, we have:

- the treatment this subject received (A, B or C are the treatments),
- an indicator (1 = yes, 0 = no) of whether the subject has diabetes,
- the subject's systolic blood pressure at baseline
- the subject's systolic blood pressure after the application of the treatment

```
# create some temporary variables
subject <- 101:115
x1 <- c("A", "B", "C", "A", "C", "A", "A", "A", NA, "B", "C", "A", "B", "C", "A", "B")
```

```

x2 <- c(1, 0, 0, 1, NA, 1, 0, 1, NA, 1, 0, 0, 1, 1, NA)
x3 <- c(120, 145, 150, NA, 155, NA, 135, NA, 115, 170, 150, 145, 140, 160, 135)
x4 <- c(105, 135, 150, 120, 135, 115, 160, 150, 130, 155, 140, 140, 150, 135, 120)

sbp_example <-
  tibble(subject, treat = factor(x1), diabetes = x2,
         sbp.before = x3, sbp.after = x4)

rm(subject, x1, x2, x3, x4) # just cleaning up

sbp_example

```

```

# A tibble: 15 x 5
  subject treat diabetes sbp.before sbp.after
  <int> <fct>    <dbl>      <dbl>     <dbl>
1     101 A          1        120       105
2     102 B          0        145       135
3     103 C          0        150       150
4     104 A          1        NA        120
5     105 C          NA       155       135
6     106 A          1        NA        115
7     107 A          0        135       160
8     108 <NA>        1        NA        150
9     109 B          NA       115       130
10    110 C          1        170       155
11    111 A          0        150       140
12    112 B          0        145       140
13    113 C          1        140       150
14    114 A          1        160       135
15    115 B          NA       135       120

```

8.3 Identifying missingness with naniar functions

The `naniar` package has many useful functions.

1. How many missing values do we have, overall?

```
n_miss(sbp_example)
```

```
[1] 7
```

2. How many variables have missing values, overall?

```
n_var_miss(sbp_example)
```

```
[1] 3
```

3. Which variables contain missing values?

```
miss_var_which(sbp_example)
```

```
[1] "treat"      "diabetes"    "sbp.before"
```

4. How many missing values do we have in each variable?

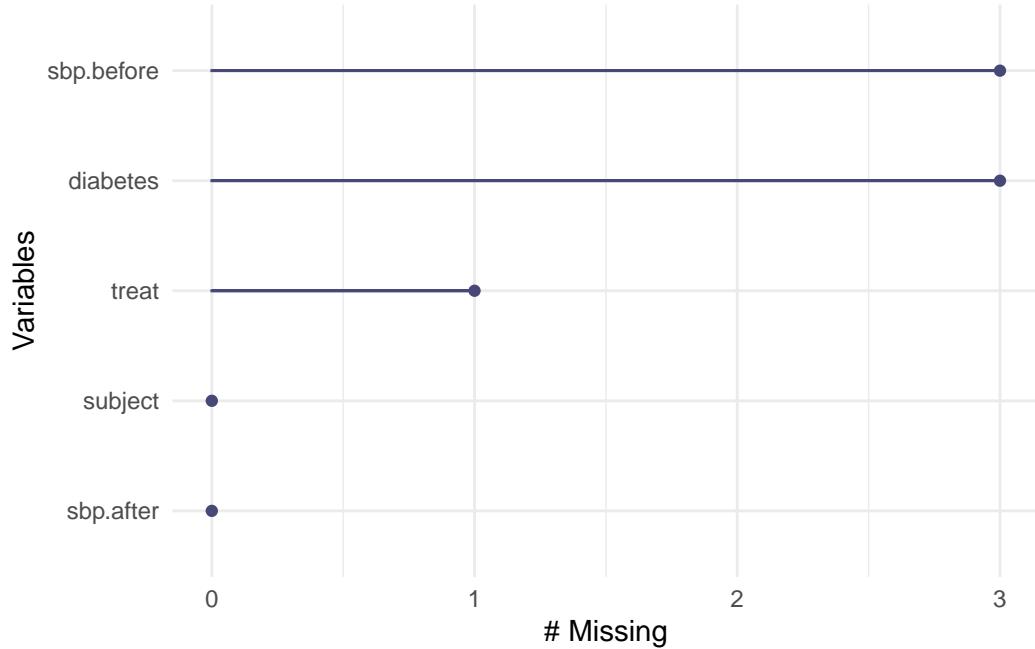
```
miss_var_summary(sbp_example)
```

```
# A tibble: 5 x 3
  variable   n_miss pct_miss
  <chr>       <int>    <dbl>
1 diabetes      3     20
2 sbp.before    3     20
3 treat         1     6.67
4 subject       0      0
5 sbp.after     0      0
```

We are missing one `treat`, 3 `diabetes` and 3 `sbp.before` values.

5. Can we plot missingness, by variable?

```
gg_miss_var(sbp_example)
```



6. How many of the cases (rows) have missing values?

```
n_case_miss(sbp_example)
```

[1] 6

7. How many cases have complete data, with no missing values?

```
n_case_complete(sbp_example)
```

[1] 9

8. Can we tabulate missingness by case?

```
miss_case_table(sbp_example)
```

```
# A tibble: 3 x 3
  n_miss_in_case n_cases pct_cases
  <int>     <int>    <dbl>
1          0         9      60
2          1         5     33.3
3          2         1     6.67
```

9. Which cases have missing values?

```
miss_case_summary(sbp_example)
```

```
# A tibble: 15 x 3
  case n_miss pct_miss
  <int>   <int>    <dbl>
1     8       2      40
2     4       1      20
3     5       1      20
4     6       1      20
5     9       1      20
6    15       1      20
7     1       0      0
8     2       0      0
9     3       0      0
10    7       0      0
11    10      0      0
12    11      0      0
13    12      0      0
14    13      0      0
15    14      0      0
```

10. How can we identify the subjects with missing data?

```
sbp_example |> filter(!complete.cases(sbp_example))
```

```
# A tibble: 6 x 5
  subject treat diabetes sbp.before sbp.after
  <int>   <fct>    <dbl>        <dbl>      <dbl>
1     104 A          1        NA      120
2     105 C         NA      155      135
3     106 A          1        NA      115
4     108 <NA>        1        NA      150
5     109 B         NA      115      130
6     115 B         NA      135      120
```

We have nine subjects with complete data, three subjects with missing `diabetes` (only), two subjects with missing `sbp.before` (only), and 1 subject who is missing both `treat` and `sbp.before`.

8.4 Missing-data mechanisms

My source for this description of mechanisms is Chapter 25 of Gelman and Hill (2007), and that chapter is [available at this link](#).

1. **MCAR = Missingness completely at random.** A variable is missing completely at random if the probability of missingness is the same for all units, for example, if for each subject, we decide whether to collect the `diabetes` status by rolling a die and refusing to answer if a “6” shows up. If data are missing completely at random, then throwing out cases with missing data does not bias your inferences.
2. **Missingness that depends only on observed predictors.** A more general assumption, called **missing at random** or **MAR**, is that the probability a variable is missing depends only on available information. Here, we would have to be willing to assume that the probability of nonresponse to `diabetes` depends only on the other, fully recorded variables in the data. It is often reasonable to model this process as a logistic regression, where the outcome variable equals 1 for observed cases and 0 for missing. When an outcome variable is missing at random, it is acceptable to exclude the missing cases (that is, to treat them as NA), as long as the regression controls for all the variables that affect the probability of missingness.
3. **Missingness that depends on unobserved predictors.** Missingness is no longer “at random” if it depends on information that has not been recorded and this information also predicts the missing values. If a particular treatment causes discomfort, a patient is more likely to drop out of the study. This missingness is not at random (unless “discomfort” is measured and observed for all patients). If missingness is not at random, it must be explicitly modeled, or else you must accept some bias in your inferences.
4. **Missingness that depends on the missing value itself.** Finally, a particularly difficult situation arises when the probability of missingness depends on the (potentially missing) variable itself. For example, suppose that people with higher earnings are less likely to reveal them.

Essentially, situations 3 and 4 are referred to collectively as **non-random missingness**, and cause more trouble for us than 1 and 2.

8.5 Options for Dealing with Missingness

There are several available methods for dealing with missing data that are MCAR or MAR, but they basically boil down to:

- Complete Case (or Available Case) analyses
- Single Imputation
- Multiple Imputation

8.6 Complete Case (and Available Case) analyses

In **Complete Case** analyses, rows containing NA values are omitted from the data before analyses commence. This is the default approach for many statistical software packages, and may introduce unpredictable bias and fail to include some useful, often hard-won information.

- A complete case analysis can be appropriate when the number of missing observations is not large, and the missing pattern is either MCAR (missing completely at random) or MAR (missing at random.)
- Two problems arise with complete-case analysis:
 1. If the units with missing values differ systematically from the completely observed cases, this could bias the complete-case analysis.
 2. If many variables are included in a model, there may be very few complete cases, so that most of the data would be discarded for the sake of a straightforward analysis.
- A related approach is *available-case* analysis where different aspects of a problem are studied with different subsets of the data, perhaps identified on the basis of what is missing in them.

8.7 Single Imputation

In **single imputation** analyses, NA values are estimated/replaced *one time* with *one particular data value* for the purpose of obtaining more complete samples, at the expense of creating some potential bias in the eventual conclusions or obtaining slightly *less* accurate estimates than would be available if there were no missing values in the data.

- A single imputation can be just a replacement with the mean or median (for a quantity) or the mode (for a categorical variable.) However, such an approach, though easy to understand, underestimates variance and ignores the relationship of missing values to other variables.
- Single imputation can also be done using a variety of models to try to capture information about the NA values that are available in other variables within the data set.
- The **simputation** package can help us execute single imputations using a wide variety of techniques, within the pipe approach used by the **tidyverse**. Another approach I have used in the past is the **mice** package, which can also perform single imputations.

8.8 Multiple Imputation

Multiple imputation, where NA values are repeatedly estimated/replaced with multiple data values, for the purpose of obtaining more complete samples *and* capturing details of the

variation inherent in the fact that the data have missingness, so as to obtain *more* accurate estimates than are possible with single imputation.

- We'll postpone further discussion of multiple imputation to later in the semester.

8.9 Building a Complete Case Analysis

We can drop all of the missing values from a data set with `drop_na` or with `na.omit` or by filtering for `complete.cases`. Any of these approaches produces the same result - a new data set with 9 rows (after dropping the six subjects with any NA values) and 5 columns.

```
cc.1 <- na.omit(sbp_example)
cc.2 <- sbp_example |> drop_na()
cc.3 <- sbp_example |> filter(complete.cases(sbp_example))
```

8.10 Single Imputation with the Mean or Mode

The most straightforward approach to single imputation is to impute a single summary of the variable, such as the mean, median or mode.

```
mosaic::favstats(~ sbp.before, data = sbp_example)

min   Q1 median      Q3 max      mean      sd n missing
115 135     145 151.25 170 143.3333 15.71527 12       3

sbp_example |> tabyl(diabetes, treat) |>
  adorn_totals(where = c("row", "col"))

diabetes A B C NA_ Total
      0 2 2 1    0    5
      1 4 0 2    1    7
<NA> 0 2 1    0    3
Total 6 4 4    1   15
```

Here, suppose we decide to impute

- `sbp.before` with the mean (143.3) among non-missing values,
- `diabetes` with its more common value, 1, and

- `treat` with its more common value, or mode (A)

```
si.1 <- sbp_example |>
  replace_na(list(sbp.before = 143.33,
                  diabetes = 1,
                  treat = "A"))

si.1
```

	subject	treat	diabetes	sbp.before	sbp.after
	<int>	<fct>	<dbl>	<dbl>	<dbl>
1	101	A	1	120	105
2	102	B	0	145	135
3	103	C	0	150	150
4	104	A	1	143.	120
5	105	C	1	155	135
6	106	A	1	143.	115
7	107	A	0	135	160
8	108	A	1	143.	150
9	109	B	1	115	130
10	110	C	1	170	155
11	111	A	0	150	140
12	112	B	0	145	140
13	113	C	1	140	150
14	114	A	1	160	135
15	115	B	1	135	120

8.11 Doing Single Imputation with `simputation`

Single imputation is a potentially appropriate method when missingness can be assumed to be either completely at random (MCAR) or dependent only on observed predictors (MAR). We'll use the `simputation` package to accomplish it.

- The `simputation` vignette is available at <https://cran.r-project.org/web/packages/simputation/vignettes/>
- The `simputation` reference manual is available at <https://cran.r-project.org/web/packages/simputation/s>

Suppose we wanted to use:

- a robust linear model to predict `sbp.before` missing values, on the basis of `sbp.after` and `diabetes` status, and

- a predictive mean matching approach (which, unlike the robust linear model, will ensure that only values of `diabetes` that we've seen before will be imputed) to predict `diabetes` status, on the basis of `sbp.after`, and
- a decision tree approach to predict `treat` status, using all other variables in the data

```
set.seed(50001)

imp.2 <- sbp_example |>
  impute_rlm(sbp.before ~ sbp.after + diabetes) |>
  impute_pmm(diabetes ~ sbp.after) |>
  impute_cart(treat ~ .)
```

```
imp.2
```

```
# A tibble: 15 x 5
  subject treat diabetes sbp.before sbp.after
* <int> <fct>     <dbl>      <dbl>      <dbl>
 1     101 A          1        120       105
 2     102 B          0        145       135
 3     103 C          0        150       150
 4     104 A          1        139.      120
 5     105 C          1        155       135
 6     106 A          1        136.      115
 7     107 A          0        135       160
 8     108 A          1        155.      150
 9     109 B          1        115       130
10    110 C          1        170       155
11    111 A          0        150       140
12    112 B          0        145       140
13    113 C          1        140       150
14    114 A          1        160       135
15    115 B          1        135       120
```

Details on the many available methods in `simputation` are provided [in its manual](#). These include:

- `impute_cart` uses a Classification and Regression Tree approach for numerical or categorical data. There is also an `impute_rf` command which uses Random Forests for imputation.
- `impute_pmm` is one of several “hot deck” options for imputation, this one is predictive mean matching, which can be used with numeric data (only). Missing values are first imputed using a predictive model. Next, these predictions are replaced with the observed

values which are nearest to the prediction. Other imputation options in this group include random hot deck, sequential hot deck and k-nearest neighbor imputation.

- `impute_rlm` is one of several regression imputation methods, including linear models, robust linear models (which use what is called M-estimation to impute numerical variables) and lasso/elastic net/ridge regression models.

The `simputation` package can also do EM-based multivariate imputation, and multivariate random forest imputation, and several other approaches.

9 National Youth Fitness Survey

9.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(janitor)
library(knitr)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

We also use functions from the `Hmisc` and `mosaic` packages in this chapter, but do not load the whole packages.

9.2 What is the NHANES NYFS?

The `nnyfs.csv` and the `nnyfs.Rds` data files were built by Professor Love using data from the [2012 National Youth Fitness Survey](#).

The NHANES National Youth Fitness Survey (NNYFS) was conducted in 2012 to collect data on physical activity and fitness levels in order to provide an evaluation of the health and fitness of children in the U.S. ages 3 to 15. The NNYFS collected data on physical activity and fitness levels of our youth through interviews and fitness tests.

In the `nnyfs` data file (either `.csv` or `.Rds`), I'm only providing a modest fraction of the available information. More on the NNYFS (including information I'm not using) is available at <https://wwwn.cdc.gov/nchs/nhanes/search/nnyfs12.aspx>.

The data elements I'm using fall into four main groups, or components:

- Demographics
- Dietary
- Examination and

- Questionnaire

What I did was merge a few elements from each of the available components of the NHANES National Youth Fitness Survey, reformulated (and in some cases simplified) some variables, and restricted the sample to kids who had completed elements of each of the four components.

9.3 The Variables included in nnyfs

This section tells you where the data come from, and briefly describe what is collected.

9.3.1 From the NNYFS Demographic Component

All of these come from the Y_DEMO file.

In nnyfs	In Y_DEMO	Description
SEQN	SEQN	Subject ID, connects all of the files
sex	RIAGENDR	Really, this is sex, not gender
age_child	RIDAGEYR	Age in years at screening
race_eth	RIDRETH1	Race/Hispanic origin (collapsed to 4 levels)
educ_child	DMDEDUC3	Education Level (for children ages 6-15). 0 = Kindergarten, 9 = Ninth grade or higher
language	SIALANG	Language in which the interview was conducted
sampling_wt	WTMEC	Full-sample MEC exam weight (for inference)
income_pov	INDFMPIR	Ratio of family income to poverty (ceiling is 5.0)
age_adult	DMDHRAGE	Age of adult who brought child to interview
educ_adult	DMDHREDU	Education level of adult who brought child

9.3.2 From the NNYFS Dietary Component

From the Y_DR1TOT file, we have a number of variables related to the child's diet, with the following summaries mostly describing consumption "yesterday" in a dietary recall questionnaire.

In nnyfs	In Y_DR1TOT	Description
respondent	DR1MNRSR	who responded to interview (child, Mom, someone else)
salt_used	DBQ095Z	uses salt, lite salt or salt substitute at the table
energy	DR1TKCAL	energy consumed (kcal)
protein	DR1TPROT	protein consumed (g)

In nnyfs	In Y_DR1TOT	Description
sugar	DR1TSUGR	total sugar consumed (g)
fat	DR1TTFAT	total fat consumed (g)
diet_yesterday	DR1_300	compare food consumed yesterday to usual amount
water	DR1_320Z	total plain water drank (g)

9.3.3 From the NNYFS Examination Component

From the Y_BMX file of Body Measures:

In nnyfs	In Y_BMX	Description
height	BMXHT	standing height (cm)
weight	BMXWT	weight (kg)
bmi	BMXBMI	body mass index (kg/m^2)
bmi_cat	BMDBMIC	BMI category (4 levels)
arm_length	BMXARML	Upper arm length (cm)
waist	BMXWAIST	Waist circumference (cm)
arm_circ	BMXARMC	Arm circumference (cm)
calf_circ	BMXCALF	Maximal calf circumference (cm)
calf_skinfold	BMXCALFF	Calf skinfold (mm)
triceps_skinfold	BMXTRI	Triceps skinfold (mm)
subscapular_skinfold	BMXSUB	Subscapular skinfold (mm)

From the Y_PLX file of Plank test results:

In nnyfs	In Y_PLX	Description
plank_time	MPXPLANK	# of seconds plank position is held

9.3.4 From the NNYFS Questionnaire Component

From the Y_PAQ file of Physical Activity questions:

In nnyfs	In Y_PAQ	Description
active_days	PAQ706	Days physically active (≥ 60 min.) in past week
tv_hours	PAQ710	Average hours watching TV/videos past 30d
computer_hours	PAQ715	Average hours on computer past 30d
physical_last_week	PAQ722	Any physical activity outside of school past week

In nnyfs	In Y_PAQ	Description
enjoy_recess	PAQ750	Enjoy participating in PE/recess

From the Y_DBQ file of Diet Behavior and Nutrition questions:

In nnyfs	In Y_DBQ	Description
meals_out	DBD895	# meals not home-prepared in past 7 days

From the Y_HIQ file of Health Insurance questions:

In nnyfs	In Y_HIQ	Description
insured	HIQ011	Covered by Health Insurance?
insurance	HIQ031	Type of Health Insurance coverage

From the Y_HUQ file of Access to Care questions:

In nnyfs	In Y_HUQ	Description
phys_health	HUQ010	General health condition (Excellent - Poor)
access_to_care	HUQ030	Routine place to get care?
care_source	HUQ040	Type of place most often goes to for care

From the Y_MCQ file of Medical Conditions questions:

In nnyfs	In Y_MCQ	Description
asthma_ever	MCQ010	Ever told you have asthma?
asthma_now	MCQ035	Still have asthma?

From the Y_RXQ_RX file of Prescription Medication questions:

In nnyfs	In Y_RXQ_RX	Description
med_use	RXDUSE	Taken prescription medication in last month?
med_count	RXDCOUNT	# of prescription meds taken in past month

9.4 Looking over A Few Variables

Now, I'll take a look at the `nnyfs` data, which I've made available in a comma-separated version (`nnyfs.csv`), if you prefer, as well as in an R data set (`nnyfs.Rds`) which loads a bit faster. After loading the file, let's get a handle on its size and contents. In my R Project for these notes, the data are contained in a separate `data` subdirectory.

```
nnyfs <- readRDS("data/nnyfs.Rds")  
  
## size of the tibble  
dim(nnyfs)
```

```
[1] 1518 45
```

There are 1518 rows (subjects) and 45 columns (variables), by which I mean that there are 1518 kids in the `nnyfs` data frame, and we have 45 pieces of information on each subject. So, what do we have, exactly?

```
nnyfs # this is a tibble, has some nice features in a print-out like this  
  
# A tibble: 1,518 x 45  
# ... with 1,508 more rows, 35 more variables: respondent <fct>,  
#   salt_used <fct>, energy <dbl>, protein <dbl>, sugar <dbl>, fat <dbl>,  
#   diet_yesterday <fct>, water <dbl>, plank_time <dbl>, height <dbl>,  
#   weight <dbl>, bmi <dbl>, bmi_cat <fct>, arm_length <dbl>, waist <dbl>,  
#   arm_circ <dbl>, calf_circ <dbl>, calf_skinfold <dbl>,  
#   triceps_skinfold <dbl>, subscapular_skinfold <dbl>, active_days <dbl>,  
#   tv_hours <dbl>, computer_hours <dbl>, physical_last_week <fct>, ...  
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Tibbles are a modern reimagining of the main way in which people have stored data in R, called a data frame. Tibbles were developed to keep what time has proven to be effective, and throwing out what is not. We can learn something about the structure of the tibble from such functions as `str` or `glimpse`.

```
str(nnyfs)
```

```
tibble [1,518 x 45] (S3: tbl_df/tbl/data.frame)
$ SEQN           : num [1:1518] 71917 71918 71919 71920 71921 ...
$ sex            : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 2 2 1 2 2 ...
$ age_child      : num [1:1518] 15 8 14 15 3 12 12 8 7 8 ...
$ race_eth       : Factor w/ 4 levels "1_Hispanic","2_White Non-Hispanic",...: 3 3 2 2 ...
$ educ_child     : num [1:1518] 9 2 8 8 NA 6 5 2 0 2 ...
$ language        : Factor w/ 2 levels "English","Spanish": 1 1 1 1 1 1 1 1 1 1 ...
$ sampling_wt    : num [1:1518] 28299 15127 29977 80652 55592 ...
$ income_pov     : num [1:1518] 0.21 5 5 0.87 4.34 5 5 2.74 0.46 1.57 ...
$ age_adult      : num [1:1518] 46 46 42 53 31 42 39 31 45 56 ...
$ educ_adult     : Factor w/ 5 levels "1_Less than 9th Grade",...: 2 3 5 3 3 4 2 3 2 3 ...
$ respondent     : Factor w/ 3 levels "Child","Mom",...: 1 2 1 1 2 1 1 1 2 1 ...
$ salt_used       : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 2 2 1 2 ...
$ energy          : num [1:1518] 2844 1725 2304 1114 1655 ...
$ protein         : num [1:1518] 169.1 55.2 199.3 14 50.6 ...
$ sugar           : num [1:1518] 128.2 118.7 81.4 119.2 90.3 ...
$ fat              : num [1:1518] 127.9 63.7 86.1 36 53.3 ...
$ diet_yesterday  : Factor w/ 3 levels "1_Much more than usual",...: 2 2 2 2 2 2 1 2 2 3 ...
$ water            : num [1:1518] 607 178 503 859 148 ...
$ plank_time      : num [1:1518] NA 45 121 45 11 107 127 44 184 58 ...
$ height          : num [1:1518] NA 131.6 172 167.1 90.2 ...
$ weight          : num [1:1518] NA 38.6 58.7 92.5 12.4 66.4 56.7 22.2 20.9 28.3 ...
$ bmi              : num [1:1518] NA 22.3 19.8 33.1 15.2 25.9 22.5 14.4 15.9 17 ...
$ bmi_cat         : Factor w/ 4 levels "1_Underweight",...: NA 4 2 4 2 4 3 2 2 2 ...
$ arm_length      : num [1:1518] NA 27.7 38.4 35.9 18.3 34.2 33 26.5 24.2 26 ...
$ waist            : num [1:1518] NA 71.9 79.4 96.4 46.8 90 72.3 56.1 54.5 59.7 ...
$ arm_circ        : num [1:1518] NA 25.4 26 37.9 15.1 29.5 27.9 17.6 17.7 19.9 ...
$ calf_circ       : num [1:1518] NA 32.3 35.3 46.8 19.4 36.9 36.8 24 24.3 27.3 ...
$ calf_s Skinfold: num [1:1518] NA 22 18.4 NA 8.4 22 18.3 7 7.2 8.2 ...
$ triceps_s Skinfold: num [1:1518] NA 19.9 15 20.6 8.6 22.8 20.5 12.9 6.9 8.8 ...
$ subscapular_s Skinfold: num [1:1518] NA 17.4 9.8 22.8 5.7 24.4 12.6 6.8 4.8 6.1 ...
$ active_days     : num [1:1518] 3 5 3 3 7 2 5 3 7 7 ...
$ tv_hours         : num [1:1518] 2 2 1 3 2 3 0 4 2 2 ...
$ computer_hours   : num [1:1518] 1 2 3 3 0 1 0 3 1 1 ...
$ physical_last_week: Factor w/ 2 levels "No","Yes": 1 1 2 2 2 2 2 2 2 2 ...
```

```

$ enjoy_recess      : Factor w/ 5 levels "1_Strongly Agree",...: 1 1 3 2 NA 2 2 NA 1 1 ...
$ meals_out         : num [1:1518] 0 2 3 2 1 1 2 1 0 2 ...
$ insured           : Factor w/ 2 levels "Has Insurance",...: 1 1 1 1 1 1 1 1 1 1 ...
$ phys_health       : Factor w/ 5 levels "1_Excellent",...: 1 3 1 3 1 1 3 1 2 1 ...
$ access_to_care    : Factor w/ 2 levels "Has Usual Care Source",...: 1 1 1 1 1 1 1 1 1 1 ...
$ care_source       : Factor w/ 6 levels "Clinic or Health Center",...: 1 2 2 2 2 2 ...
$ asthma_ever       : Factor w/ 2 levels "History of Asthma",...: 2 1 2 1 2 2 2 2 2 2 ...
$ asthma_now        : Factor w/ 2 levels "Asthma Now","No Asthma Now": 2 1 2 1 2 2 2 2 2 2 ...
$ med_use           : Factor w/ 2 levels "Had Medication",...: 2 1 2 1 2 2 2 2 2 2 ...
$ med_count         : num [1:1518] 0 1 0 2 0 0 0 0 0 0 ...
$ insurance         : Factor w/ 10 levels "Medicaid","Medicare",...: 8 8 5 8 5 5 5 5 8 1 ...

```

There are a lot of variables here. Let's run through the first few in a little detail.

9.4.1 SEQN

The first variable, `SEQN` is just a (numerical) identifying code attributable to a given subject of the survey. This is *nominal* data, which will be of little interest down the line. On some occasions, as in this case, the ID numbers are sequential, in the sense that subject 71919 was included in the data base after subject 71918, but this fact isn't particularly interesting here, because the protocol remained unchanged throughout the study.

9.4.2 sex

The second variable, `sex`, is listed as a factor variable (R uses `factor` and `character` to refer to categorical, especially non-numeric information). Here, as we can see below, we have two levels, *Female* and *Male*.

```

nnyfs |>
  tabyl(sex) |>
  adorn_totals() |>
  adorn_pct_formatting()

  sex      n percent
Female   760   50.1%
  Male   758   49.9%
  Total 1518 100.0%

```

9.4.3 age_child

The third variable, `age_child`, is the age of the child at the time of their screening to be in the study, measured in years. Note that age is a continuous concept, but the measure used here (number of full years alive) is a common discrete approach to measurement. Age, of course, has a meaningful zero point, so this can be thought of as a ratio variable; a child who is 6 is half as old as one who is 12. We can tabulate the observed values, since there are only a dozen or so.

```
nnyfs |> tabyl(age_child) |>  
  adorn_pct_formatting()
```

age_child	n	percent
3	110	7.2%
4	112	7.4%
5	114	7.5%
6	129	8.5%
7	123	8.1%
8	112	7.4%
9	99	6.5%
10	124	8.2%
11	111	7.3%
12	137	9.0%
13	119	7.8%
14	130	8.6%
15	98	6.5%

At the time of initial screening, these children should have been between 3 and 15 years of age, so things look reasonable. Since this is a meaningful quantitative variable, we may be interested in a more descriptive summary.

```
nnyfs |> select(age_child) |>  
  summary()
```

```
age_child  
Min.   : 3.000  
1st Qu.: 6.000  
Median : 9.000  
Mean   : 9.033  
3rd Qu.:12.000  
Max.   :15.000
```

These six numbers provide a nice, if incomplete, look at the ages.

- **Min.** = the minimum, or youngest age at the examination was 3 years old.
- **1st Qu.** = the first quartile (25th percentile) of the ages was 6. This means that 25 percent of the subjects were age 6 or less.
- **Median** = the second quartile (50th percentile) of the ages was 9. This is often used to describe the center of the data. Half of the subjects were age 9 or less.
- **3rd Qu.** = the third quartile (75th percentile) of the ages was 12
- **Max.** = the maximum, or oldest age at the examination was 15 years.

We could get the standard deviation and a count of missing and non-missing observations with `favstats` from the `mosaic` package.

```
mosaic::favstats(~ age_child, data = nnyfs) |>  
  kable(digits = 1)
```

Registered S3 method overwritten by 'mosaic':

```
method                  from  
fortify.SpatialPolygonsDataFrame ggplot2
```

min	Q1	median	Q3	max	mean	sd	n	missing
3	6	9	12	15	9	3.7	1518	0

9.4.4 race_eth

The fourth variable in the data set is `race_eth`, which is a multi-categorical variable describing the child's race and ethnicity.

```
nnyfs |> tabyl(race_eth) |>  
  adorn_pct_formatting() |>  
  knitr::kable()
```

race_eth	n	percent
1_Hispanic	450	29.6%
2_White Non-Hispanic	610	40.2%
3_Black Non-Hispanic	338	22.3%
4_Other Race/Ethnicity	120	7.9%

And now, we get the idea of looking at whether our numerical summaries of the children's ages varies by their race/ethnicity...

```
mosaic::favstats(age_child ~ race_eth, data = nnyfs)
```

	race_eth	min	Q1	median	Q3	max	mean	sd	n	missing
1	1_Hispanic	3	5.25	9.0	12	15	8.793333	3.733846	450	0
2	2_White Non-Hispanic	3	6.00	9.0	12	15	9.137705	3.804421	610	0
3	3_Black Non-Hispanic	3	6.00	9.0	12	15	9.038462	3.576423	338	0
4	4_Other Race/Ethnicity	3	7.00	9.5	12	15	9.383333	3.427970	120	0

9.4.5 income_pov

Skipping down a bit, let's look at the family income as a multiple of the poverty level. Here's the summary.

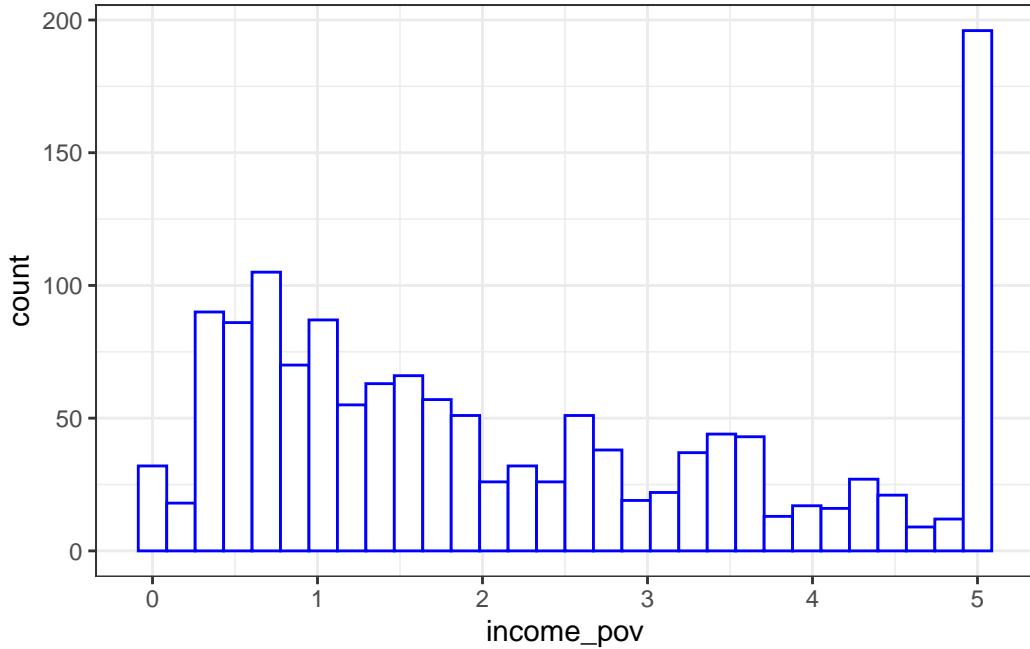
```
nnysfs |> select(income_pov) |> summary()
```

```
income_pov
Min.    :0.000
1st Qu.:0.870
Median  :1.740
Mean    :2.242
3rd Qu.:3.520
Max.    :5.000
NA's    :89
```

We see there is some missing data here. Let's ignore that for the moment and concentrate on interpreting the results for the children with actual data. We should start with a picture.

```
ggplot(nnysfs, aes(x = income_pov)) +
  geom_histogram(bins = 30, fill = "white", col = "blue")
```

```
Warning: Removed 89 rows containing non-finite values (stat_bin).
```



The histogram shows us that the values are truncated at 5, so that children whose actual family income is above 5 times the poverty line are listed as 5. We also see a message reminding us that some of the data are missing for this variable.

Is there a relationship between `income_pov` and `race_eth` in these data?

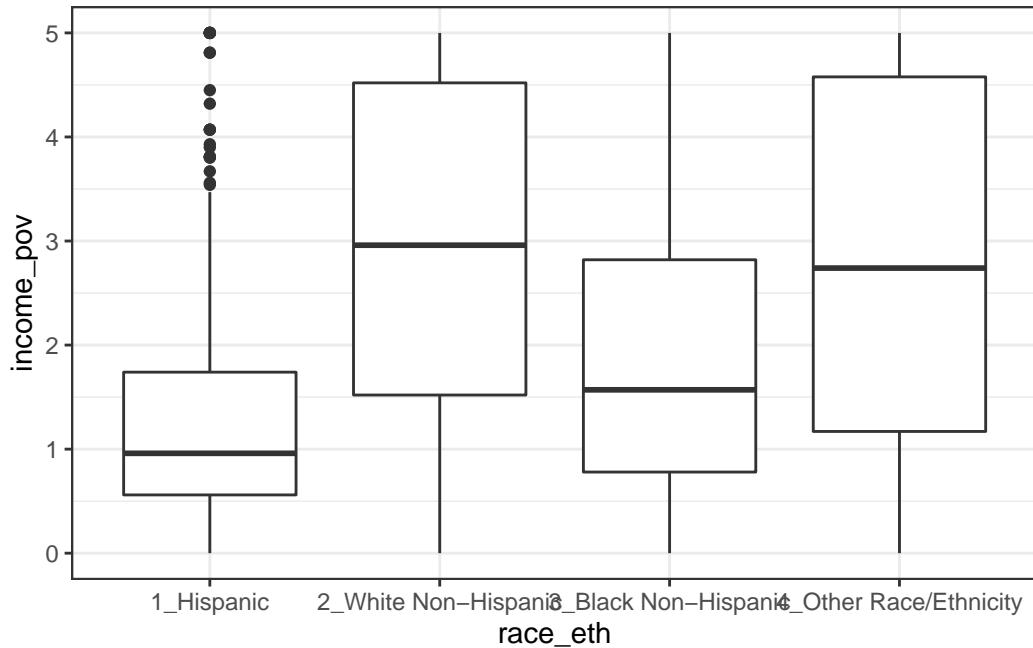
```
mosaic::favstats(income_pov ~ race_eth, data = nnyfs) |>
  kable(digits = 1)
```

race_eth	min	Q1	median	Q3	max	mean	sd	n	missing
1_Hispanic	0	0.6	1.0	1.7	5	1.3	1.1	409	41
2_White Non-Hispanic	0	1.5	3.0	4.5	5	2.9	1.6	588	22
3_Black Non-Hispanic	0	0.8	1.6	2.8	5	2.0	1.5	328	10
4_Other	0	1.2	2.7	4.6	5	2.8	1.7	104	16
Race/Ethnicity									

This deserves a picture. Let's try a boxplot.

```
ggplot(nnyfs, aes(x = race_eth, y = income_pov)) +
  geom_boxplot()
```

Warning: Removed 89 rows containing non-finite values (stat_boxplot).



9.4.6 bmi

Moving into the body measurement data, `bmi` is the body-mass index of the child. The BMI is a person's weight in kilograms divided by his or her height in meters squared. Symbolically, $BMI = \text{weight in kg} / (\text{height in m})^2$. This is a continuous concept, measured to as many decimal places as you like, and it has a meaningful zero point, so it's a ratio variable.

```
nyfs |> select(bmi) |> summary()
```

```
bmi
Min.   :11.90
1st Qu.:15.90
Median  :18.10
Mean    :19.63
3rd Qu.:21.90
Max.   :48.30
NA's   :4
```

Why would a table of these BMI values not be a great idea, for these data? A hint is that R represents this variable as `num` or numeric in its depiction of the data structure, and this implies that R has some decimal values stored. Here, I'll use the `head()` function and the `tail()` function to show the first few and the last few values of what would prove to be a very long table of `bmi` values.

```
nnyfs |> tabyl(bmi) |>  
  adorn_pct_formatting() |>  
  head()
```

bmi	n	percent	valid_percent
11.9	1	0.1%	0.1%
12.6	1	0.1%	0.1%
12.7	1	0.1%	0.1%
12.9	1	0.1%	0.1%
13.0	2	0.1%	0.1%
13.1	1	0.1%	0.1%

```
nnyfs |> tabyl(bmi) |>  
  adorn_pct_formatting() |>  
  tail()
```

bmi	n	percent	valid_percent
42.8	1	0.1%	0.1%
43.0	1	0.1%	0.1%
46.9	1	0.1%	0.1%
48.2	1	0.1%	0.1%
48.3	1	0.1%	0.1%
NA	4	0.3%	-

9.4.7 bmi_cat

Next I'll look at the `bmi_cat` information. This is a four-category ordinal variable, which divides the sample according to BMI into four groups. The BMI categories use sex-specific 2000 BMI-for-age (in months) growth charts prepared by the Centers for Disease Control for the US. We can get the breakdown from a table of the variable's values.

```
nnyfs |> tabyl(bmi_cat) |> adorn_pct_formatting()
```

	bmi_cat	n	percent	valid_percent
1_Underweight	41	2.7%	2.7%	
2_Normal	920	60.6%	60.8%	
3_Overweight	258	17.0%	17.0%	
4_Obese	295	19.4%	19.5%	
<NA>	4	0.3%	-	

In terms of percentiles by age and sex from the growth charts, the meanings of the categories are:

- Underweight ($\text{BMI} < 5\text{th percentile}$)
- Normal weight ($\text{BMI } 5\text{th to } < 85\text{th percentile}$)
- Overweight ($\text{BMI } 85\text{th to } < 95\text{th percentile}$)
- Obese ($\text{BMI} \geq 95\text{th percentile}$)

Note how I've used labels in the `bmi_cat` variable that include a number at the start so that the table results are sorted in a rational way. R sorts tables alphabetically, in general. We'll use the `forcats` package to work with categorical variables that we store as *factors* eventually, but for now, we'll keep things relatively simple.

Note that the `bmi_cat` data don't completely separate out the raw `bmi` data, because the calculation of percentiles requires different tables for each combination of `age` and `sex`.

```
mosaic::favstats(bmi ~ bmi_cat, data = nnyfs) |>
  kable(digits = 1)
```

bmi_cat	min	Q1	median	Q3	max	mean	sd	n	missing
1_Underweight	11.9	13.4	13.7	15.0	16.5	14.1	1.1	41	0
2_Normal	13.5	15.4	16.5	18.7	24.0	17.2	2.3	920	0
3_Overweight	16.9	18.3	21.4	23.4	27.9	21.2	2.9	258	0
4_Obese	17.9	22.3	26.2	30.2	48.3	26.7	5.7	295	0

9.4.8 waist

Let's also look briefly at `waist`, which is the circumference of the child's waist, in centimeters. Again, this is a numeric variable, so perhaps we'll stick to the simple summary, rather than obtaining a table of observed values.

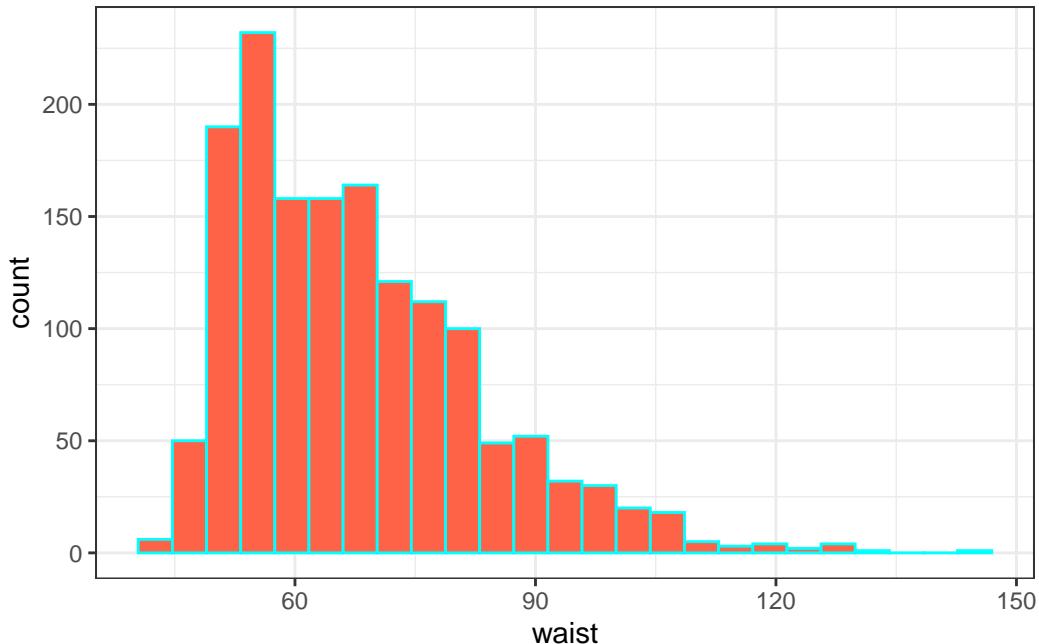
```
mosaic::favstats(~ waist, data = nnyfs)
```

min	Q1	median	Q3	max	mean	sd	n	missing
42.5	55.6	64.8	76.6	144.7	67.70536	15.19809	1512	6

Here's a histogram of the waist circumference data.

```
ggplot(nnyfs, aes(x = waist)) +
  geom_histogram(bins = 25, fill = "tomato", color = "cyan")
```

Warning: Removed 6 rows containing non-finite values (stat_bin).



9.4.9 triceps_skinfold

The last variable I'll look at for now is `triceps_skinfold`, which is measured in millimeters. This is one of several common locations used for the assessment of body fat using skinfold calipers, and is a frequent part of growth assessments in children. Again, this is a numeric variable according to R.

```
mosaic::favstats(~ triceps_skinfold, data = nnyfs)
```

```

min   Q1 median Q3  max      mean        sd     n missing
4 9.1  12.4 18 38.8 14.35725 6.758825 1497       21

```

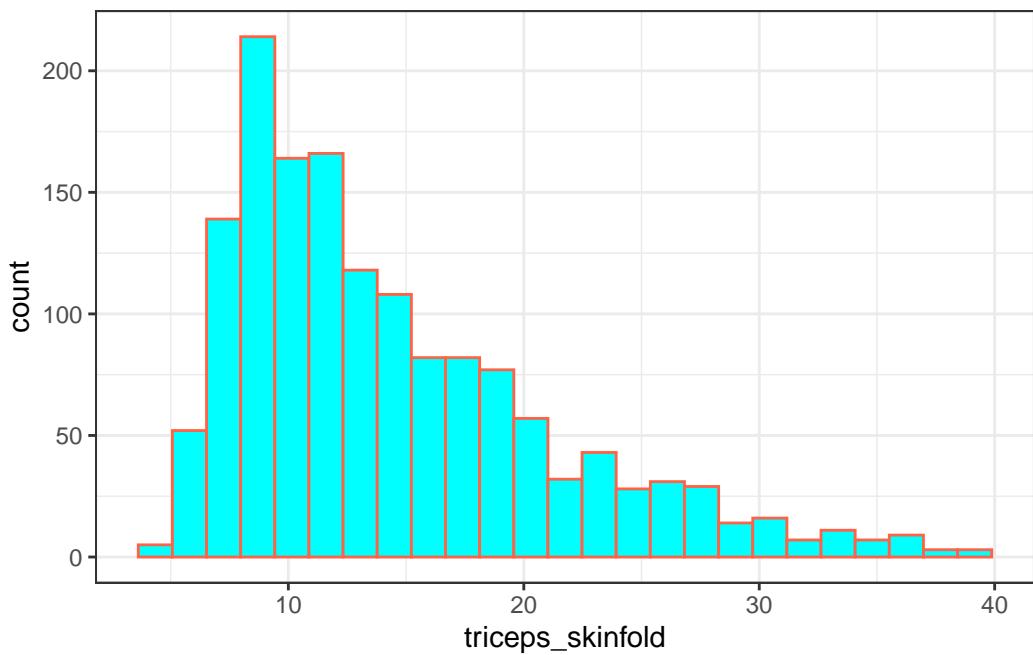
And here's a histogram of the triceps skinfold data, with the fill and color flipped from what we saw in the plot of the waist circumference data a moment ago.

```

ggplot(nnyfs, aes(x = triceps_skinfold)) +
  geom_histogram(bins = 25, fill = "cyan", color = "tomato")

```

Warning: Removed 21 rows containing non-finite values (stat_bin).



OK. We've seen a few variables, and we'll move on now to look more seriously at the data.

9.5 Additional Numeric Summaries

9.5.1 The Five Number Summary, Quantiles and IQR

The **five number summary** is most famous when used to form a box plot - it's the minimum, 25th percentile, median, 75th percentile and maximum. For numerical and integer variables,

the **summary** function produces the five number summary, plus the mean, and a count of any missing values (NA's).

```
nnyfs |>
  select(waist, energy, sugar) |>
  summary()
```

	waist	energy	sugar
Min.	: 42.50	Min. : 257	Min. : 1.00
1st Qu.	: 55.60	1st Qu.:1368	1st Qu.: 82.66
Median	: 64.80	Median :1794	Median :116.92
Mean	: 67.71	Mean :1877	Mean :124.32
3rd Qu.	: 76.60	3rd Qu.:2306	3rd Qu.:157.05
Max.	:144.70	Max. :5265	Max. :405.49
NA's	:6		

As an alternative, we can use the \$ notation to indicate the variable we wish to study inside a data set, and we can use the **fivenum** function to get the five numbers used in developing a box plot. We'll focus for a little while on the number of kilocalories consumed by each child, according to the dietary recall questionnaire. That's the **energy** variable.

```
fivenum(nnyfs$energy)
```

```
[1] 257.0 1367.0 1794.5 2306.0 5265.0
```

- As mentioned in @ref(rangeandiqr), the **inter-quartile range**, or IQR, is sometimes used as a competitor for the standard deviation. It's the difference between the 75th percentile and the 25th percentile. The 25th percentile, median, and 75th percentile are referred to as the quartiles of the data set, because, together, they split the data into quarters.

```
IQR(nnyfs$energy)
```

```
[1] 938.5
```

We can obtain **quantiles** (percentiles) as we like - here, I'm asking for the 1st and 99th:

```
quantile(nnyfs$energy, probs=c(0.01, 0.99))
```

```
1%      99%
566.85 4051.75
```

9.6 Additional Summaries from favstats

If we're focusing on a single variable, the `favstats` function in the `mosaic` package can be very helpful. Rather than calling up the entire `mosaic` library here, I'll just specify the function within the library.

```
mosaic::favstats(~ energy, data = nnyfs)

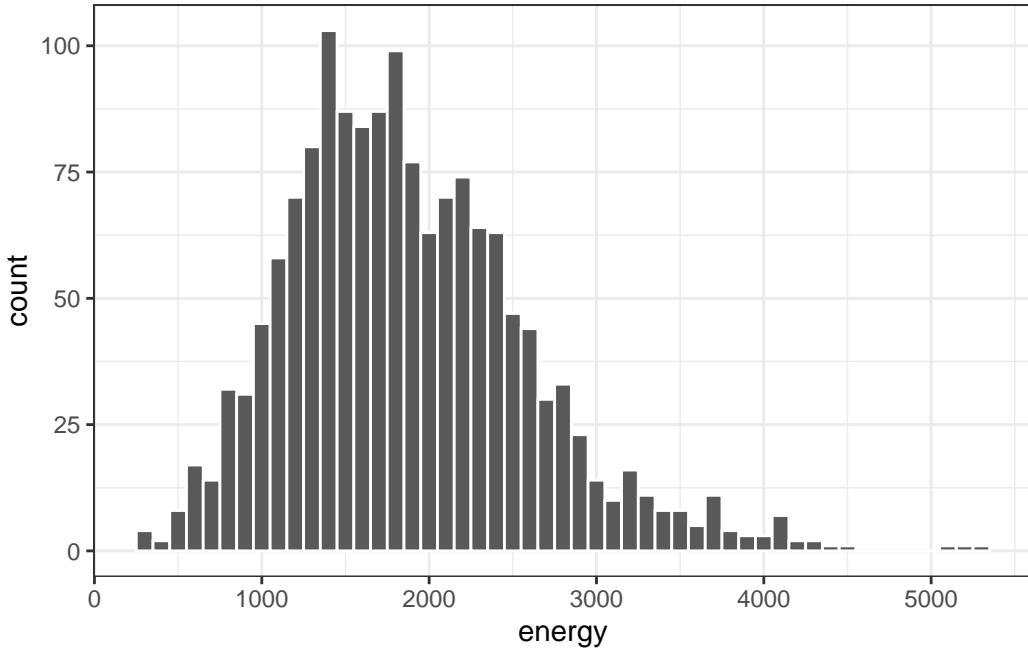
min      Q1 median      Q3   max      mean        sd      n missing
257 1367.5 1794.5 2306 5265 1877.157 722.3537 1518          0
```

This adds three useful results to the base summary - the standard deviation, the sample size and the number of missing observations.

9.7 The Histogram

Obtaining a basic **histogram** of, for example, the energy (kilocalories consumed) in the `nnyfs` data is pretty straightforward.

```
ggplot(data = nnyfs, aes(x = energy)) +
  geom_histogram(binwidth = 100, col = "white")
```



9.7.1 Freedman-Diaconis Rule to select bin width

If we like, we can suggest a particular number of cells for the histogram, instead of accepting the defaults. In this case, we have $n = 1518$ observations. The **Freedman-Diaconis rule** can be helpful here. That rule suggests that we set the bin-width to

$$h = \frac{2 * IQR}{n^{1/3}}$$

so that the number of bins is equal to the range of the data set (maximum - minimum) divided by h .

For the `energy` data in the `nnyfs` tibble, we have

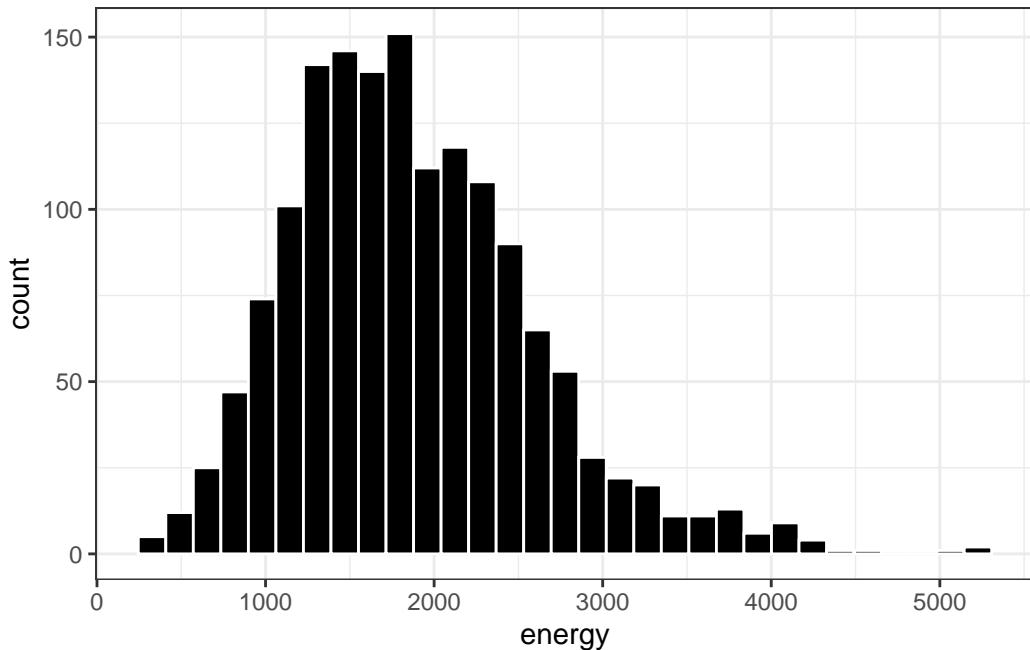
- IQR of 938.5, $n = 1518$ and range = 5008
- Thus, by the Freedman-Diaconis rule, the optimal binwidth h is 163.3203676, or, realistically, 163.
- And so the number of bins would be 30.6636586, or, realistically 31.

Here, we'll draw the graph again, using the Freedman-Diaconis rule to identify the number of bins, and also play around a bit with the fill and color of the bars.

```

bw <- 2 * IQR(nnyfs$energy) / length(nnyfs$energy)^(1/3)
ggplot(data = nnyfs, aes(x = energy)) +
  geom_histogram(binwidth=bw, color = "white", fill = "black")

```



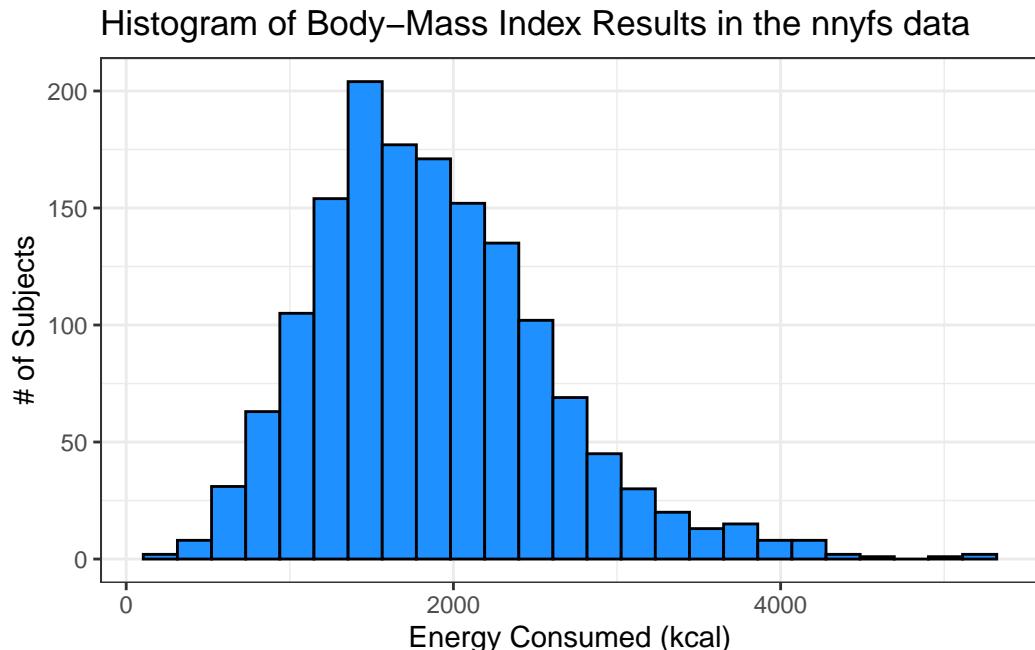
This is a nice start, but it is by no means a finished graph.

Let's improve the axis labels, add a title, and fill in the bars with a distinctive blue and use a black outline around each bar. I'll just use 25 bars, because I like how that looks in this case, and optimizing the number of bins is rarely important.

```

ggplot(data = nnyfs, aes(x = energy)) +
  geom_histogram(bins=25, color = "black", fill = "dodgerblue") +
  labs(title = "Histogram of Body-Mass Index Results in the nnyfs data",
       x = "Energy Consumed (kcal)", y = "# of Subjects")

```



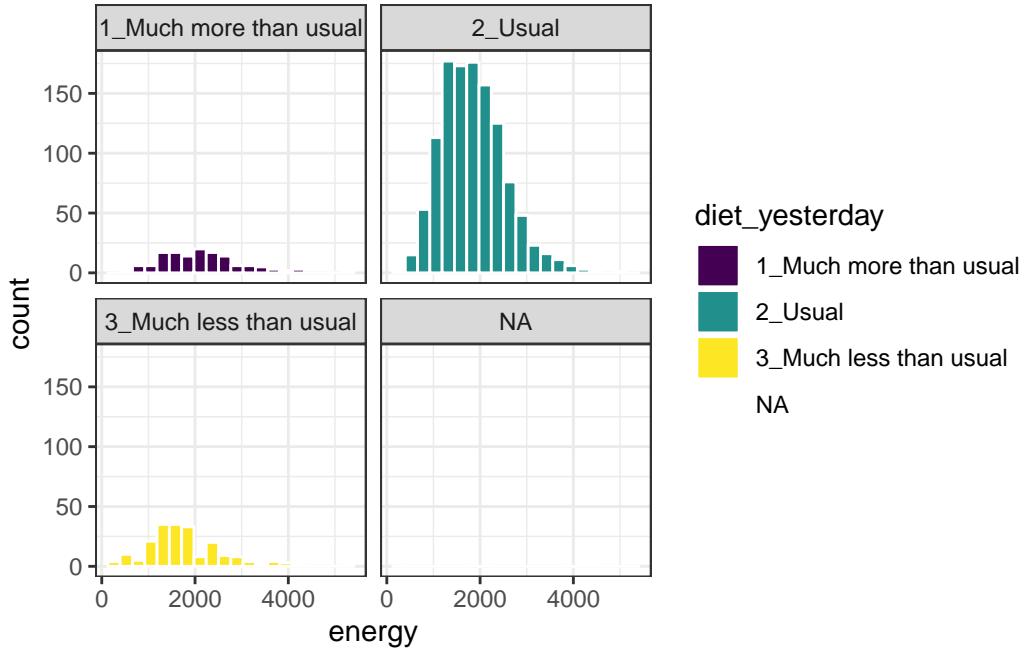
9.7.2 A Note on Colors

The simplest way to specify a color is with its name, enclosed in parentheses. My favorite list of R colors is <http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf>. In a pinch, you can usually find it by googling **Colors in R**. You can also type `colors()` in the R console to obtain a list of the names of the same 657 colors.

When using colors to make comparisons, you may be interested in using a scale that has some nice properties. The [viridis package vignette](#) describes four color scales (viridis, magma, plasma and inferno) that are designed to be colorful, robust to colorblindness and gray scale printing, and perceptually uniform, which means (as the package authors describe it) that values close to each other have similar-appearing colors and values far away from each other have more different-appearing colors, consistently across the range of values. We can apply these colors with special functions within `ggplot`.

Here's a comparison of several histograms, looking at `energy` consumed as a function of whether yesterday was typical in terms of food consumption.

```
ggplot(data = nnyfs, aes(x = energy, fill = diet_yesterday)) +
  geom_histogram(bins = 20, col = "white") +
  scale_fill_viridis_d() +
  facet_wrap(~ diet_yesterday)
```

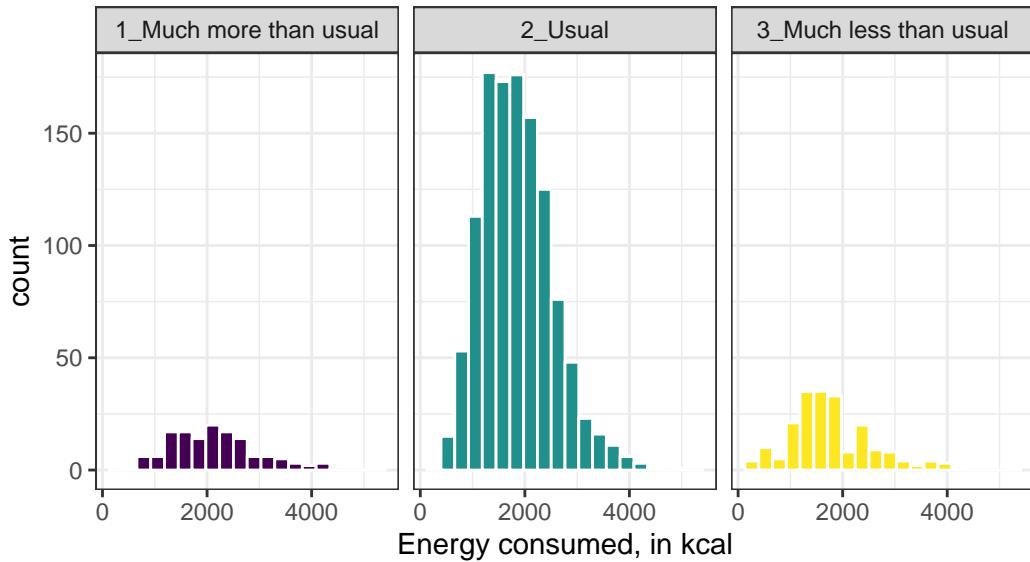


We don't really need the legend here, and perhaps we should restrict the plot to participants who responded to the `diet_yesterday` question, and put in a title and better axis labels?

```
nnyfs |> filter(!is.na(energy), !is.na(diet_yesterday)) %>%
  ggplot(data = ., aes(x = energy, fill = diet_yesterday)) +
  geom_histogram(bins = 20, col = "white") +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  facet_wrap(~ diet_yesterday) +
  labs(x = "Energy consumed, in kcal",
       title = "Energy Consumption and How Typical Was Yesterday's Eating",
       subtitle = "NHANES National Youth Fitness Survey, no survey weighting")
```

Energy Consumption and How Typical Was Yesterday's Eating

NHANES National Youth Fitness Survey, no survey weighting

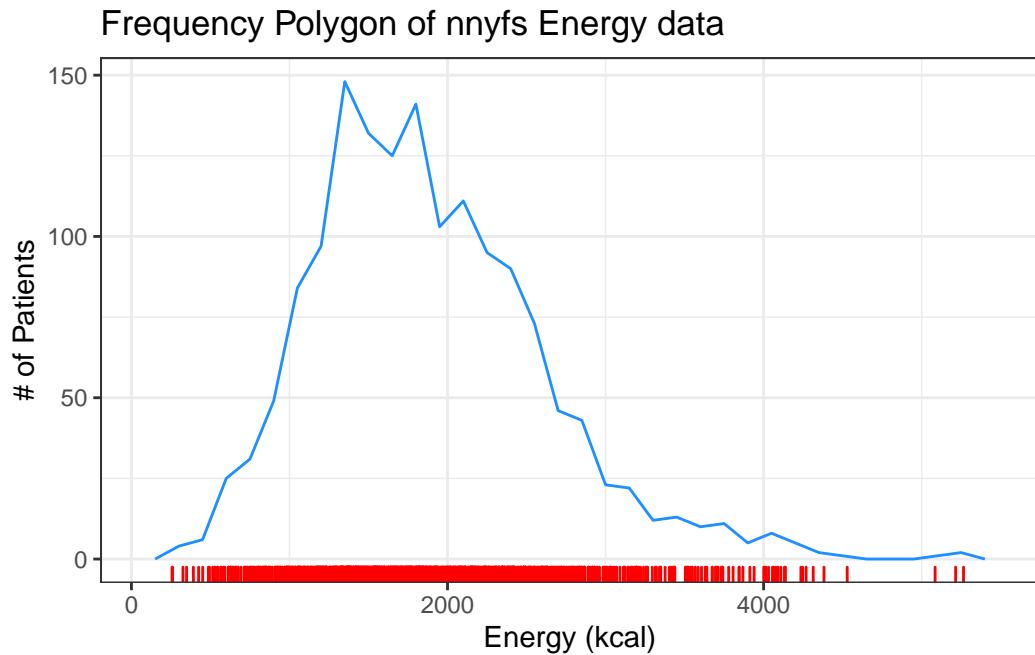


Note the use of the `%>%` pipe here. I need to write more about this.

9.8 The Frequency Polygon

As we've seen, we can also plot the distribution of a single continuous variable using the `freqpoly` geom. We can also add a *rug plot*, which places a small vertical line on the horizontal axis everywhere where an observation appears in the data.

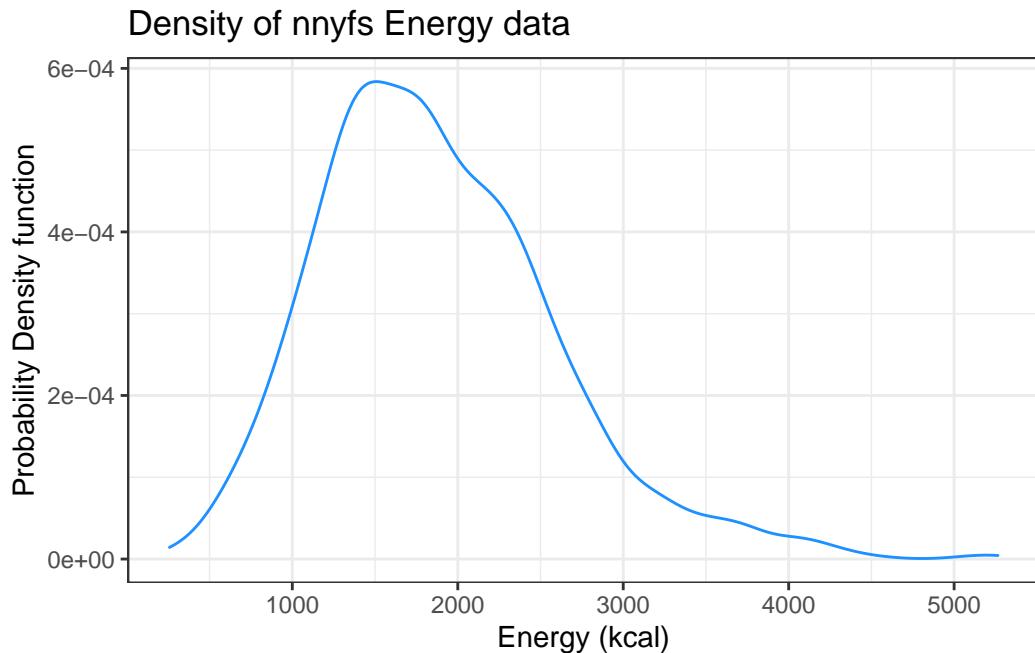
```
ggplot(data = nnyfs, aes(x = energy)) +  
  geom_freqpoly(binwidth = 150, color = "dodgerblue") +  
  geom_rug(color = "red") +  
  labs(title = "Frequency Polygon of nnyfs Energy data",  
       x = "Energy (kcal)", y = "# of Patients")
```



9.9 Plotting the Probability Density Function

We can also produce a density function, which has the effect of smoothing out the bumps in a histogram or frequency polygon, while also changing what is plotted on the y-axis.

```
ggplot(data = nnyfs, aes(x = energy)) +
  geom_density(kernel = "gaussian", color = "dodgerblue") +
  labs(title = "Density of nnyfs Energy data",
       x = "Energy (kcal)", y = "Probability Density function")
```



So, what's a density function?

- A probability density function is a function of a continuous variable, x , that represents the probability of x falling within a given range. Specifically, the integral over the interval (a,b) of the density function gives the probability that the value of x is within (a,b) .
- If you're interested in exploring more on the notion of density functions for continuous (and discrete) random variables, some nice elementary material is available at [Khan Academy](#).

9.10 The Boxplot

Sometimes, it's helpful to picture the five-number summary of the data in such a way as to get a general sense of the distribution. One approach is a **boxplot**, sometimes called a box-and-whisker plot.

9.10.1 Drawing a Boxplot for One Variable in ggplot2

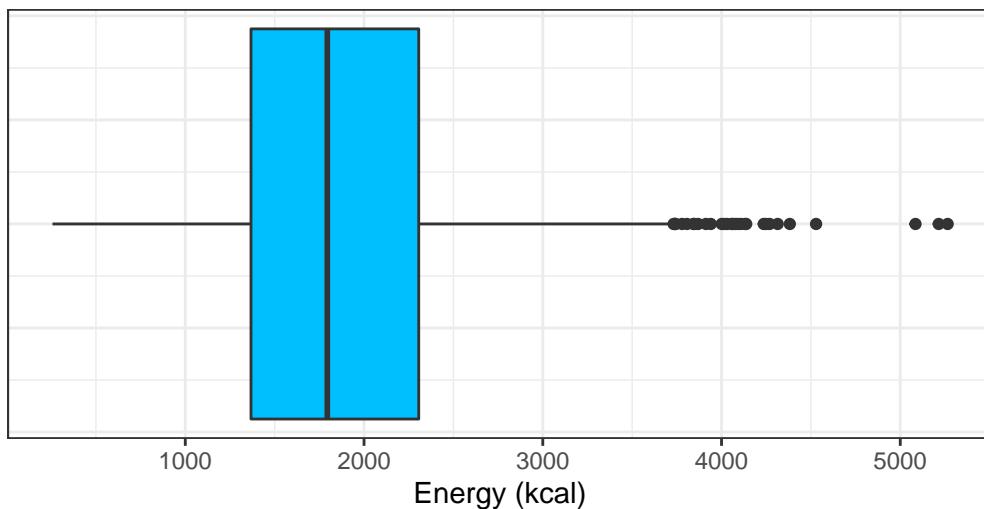
The `ggplot2` library easily handles comparison boxplots for multiple distributions, as we'll see in a moment. However, building a boxplot for a single distribution requires a little trickiness.

```

ggplot(nnyfs, aes(x = 1, y = energy)) +
  geom_boxplot(fill = "deepskyblue") +
  coord_flip() +
  labs(title = "Boxplot of Energy for kids in the NNYFS",
       y = "Energy (kcal)",
       x = "") +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())

```

Boxplot of Energy for kids in the NNYFS



9.10.2 About the Boxplot

The boxplot is another John Tukey invention.

- R draws the box (here in yellow) so that its edges of the box fall at the 25th and 75th percentiles of the data, and the thick line inside the box falls at the median (50th percentile).
- The whiskers then extend out to the largest and smallest values that are not classified by the plot as candidate *outliers*.
- An outlier is an unusual point, far from the center of a distribution.
- Note that I've used the **horizontal** option to show this boxplot in this direction. Most comparison boxplots, as we'll see below, are oriented vertically.

The boxplot's **whiskers** that are drawn from the first and third quartiles (i.e. the 25th and 75th percentiles) out to the most extreme points in the data that do not meet the standard

of “candidate outliers.” An outlier is simply a point that is far away from the center of the data - which may be due to any number of reasons, and generally indicates a need for further investigation.

Most software, including R, uses a standard proposed by Tukey which describes a “candidate outlier” as any point above the *upper fence* or below the *lower fence*. The definitions of the fences are based on the inter-quartile range (IQR).

If $IQR = 75\text{th percentile} - 25\text{th percentile}$, then the upper fence is $75\text{th percentile} + 1.5 \cdot IQR$, and the lower fence is $25\text{th percentile} - 1.5 \cdot IQR$.

So for these `energy` data,

- the upper fence is located at $2306 + 1.5(938.5) = 3713.75$
- the lower fence is located at $1367 - 1.5(938.5) = -40.75$

In this case, we see no points identified as outliers in the low part of the distribution, but quite a few identified that way on the high side. This tends to identify about 5% of the data as a candidate outlier, *if* the data follow a Normal distribution.

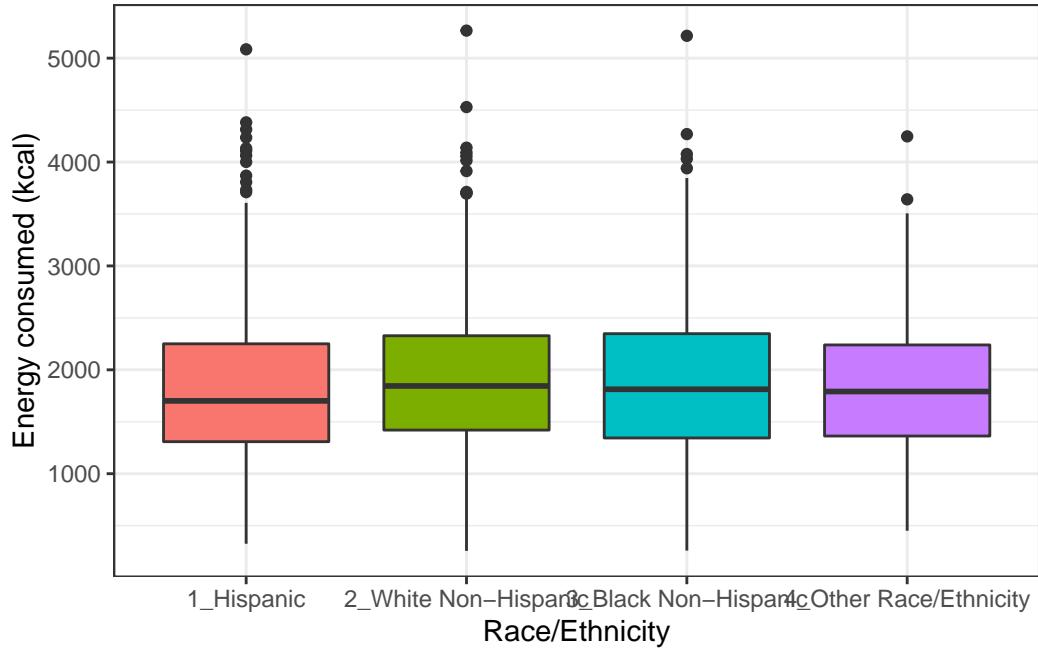
- This plot is indicating clearly that there is some asymmetry (skew) in the data, specifically right skew.
- The standard R uses is to indicate as outliers any points that are more than 1.5 inter-quartile ranges away from the edges of the box.

The horizontal orientation I’ve chosen here clarifies the relationship of direction of skew to the plot. A plot like this, with multiple outliers on the right side is indicative of a long right tail in the distribution, and hence, positive or right skew - with the mean being larger than the median. Other indications of skew include having one side of the box being substantially wider than the other, or one side of the whiskers being substantially longer than the other. More on skew later.

9.11 A Simple Comparison Boxplot

Boxplots are most often used for comparison. We can build boxplots using `ggplot2`, as well, and we’ll discuss that in detail later. For now, here’s a boxplot built to compare the `energy` results by the subject’s race/ethnicity.

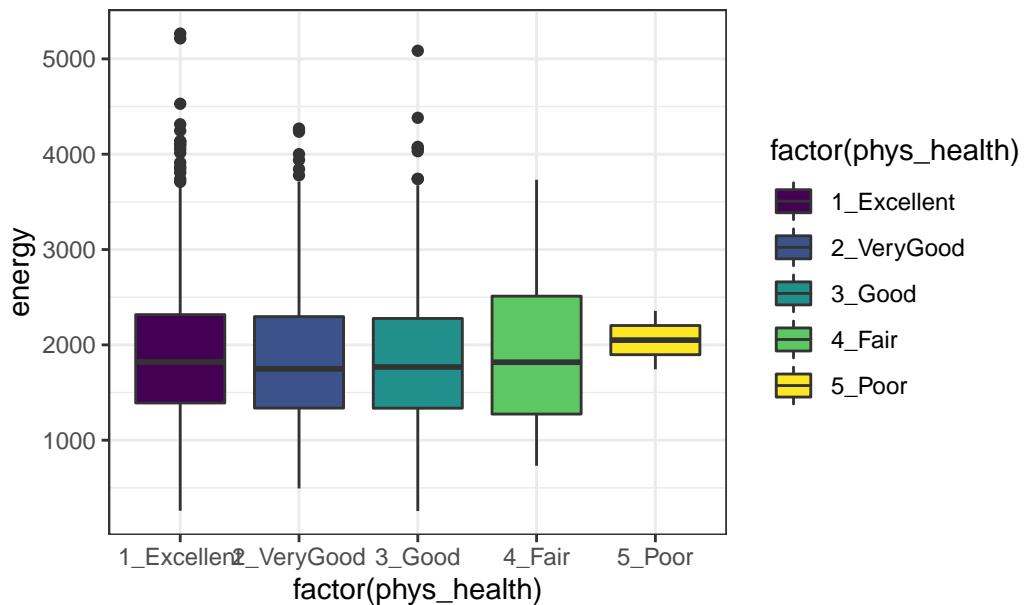
```
ggplot(nnyfs, aes(x = factor(race_eth), y = energy, fill=factor(race_eth))) +
  geom_boxplot() +
  guides(fill = "none") +
  labs(y = "Energy consumed (kcal)", x = "Race/Ethnicity")
```



Let's look at the comparison of observed energy levels across the five categories in our `phys_health` variable, now making use of the `viridis` color scheme.

```
ggplot(nnyfs, aes(x = factor(phys_health), y = energy, fill = factor(phys_health))) +
  geom_boxplot() +
  scale_fill_viridis_d() +
  labs(title = "Energy by Self-Reported Physical Health, in nnyfs data")
```

Energy by Self-Reported Physical Health, in nnyfs data

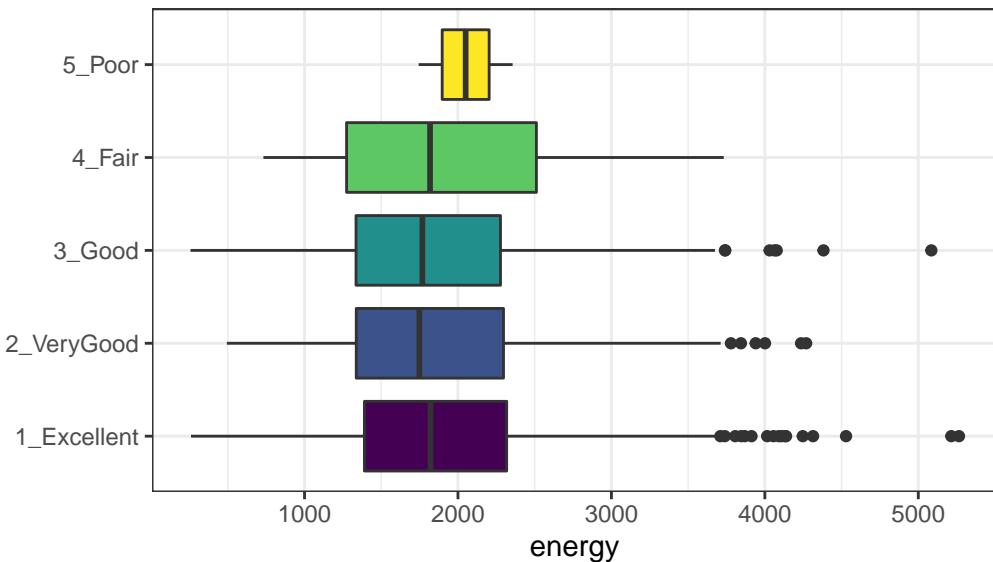


As a graph, that's not bad, but what if we want to improve it further?

Let's turn the boxes in the horizontal direction, and get rid of the perhaps unnecessary `phys_health` labels.

```
ggplot(nnyfs, aes(x = factor(phys_health), y = energy, fill = factor(phys_health))) +
  geom_boxplot() +
  scale_fill_viridis_d() +
  coord_flip() +
  guides(fill = "none") +
  labs(title = "Energy Consumed by Self-Reported Physical Health",
       subtitle = "NHANES National Youth Fitness Survey, unweighted",
       x = "")
```

Energy Consumed by Self-Reported Physical Health
NHANES National Youth Fitness Survey, unweighted



9.12 Using describe in the psych library

For additional numerical summaries, one option would be to consider using the `describe` function from the `psych` library.

```
psych::describe(nnyfs$energy)
```

```
vars     n      mean       sd median trimmed      mad min   max range skew kurtosis
X1     1 1518 1877.16 722.35 1794.5 1827.1 678.29 257 5265 5008  0.8     1.13
      se
X1 18.54
```

This package provides, in order, the following...

- `n` = the sample size
- `mean` = the sample mean
- `sd` = the sample standard deviation
- `median` = the median, or 50th percentile
- `trimmed` = mean of the middle 80% of the data
- `mad` = median absolute deviation
- `min` = minimum value in the sample

- `max` = maximum value in the sample
- `range` = max - min
- `skew` = skewness measure, described below (indicates degree of asymmetry)
- `kurtosis` = kurtosis measure, described below (indicates heaviness of tails, degree of outlier-proneness)
- `se` = standard error of the sample mean = `sd` / square root of sample size, useful in inference

9.12.1 The Trimmed Mean

The **trimmed mean** trim value in R indicates proportion of observations to be trimmed from each end of the outcome distribution before the mean is calculated. The **trimmed** value provided by the `psych::describe` package describes what this particular package calls a 20% trimmed mean (bottom and top 10% of `energy` values are removed before taking the mean - it's the mean of the middle 80% of the data.) I might call that a 10% trimmed mean in some settings, but that's just me.

```
mean(nnyfs$energy, trim=.1)
```

```
[1] 1827.1
```

9.12.2 The Median Absolute Deviation

An alternative to the IQR that is fancier, and a bit more robust, is the **median absolute deviation**, which, in large sample sizes, for data that follow a Normal distribution, will be (in expectation) equal to the standard deviation. The MAD is the median of the absolute deviations from the median, multiplied by a constant (1.4826) to yield asymptotically normal consistency.

```
mad(nnyfs$energy)
```

```
[1] 678.2895
```

9.13 Assessing Skew

A relatively common idea is to assess **skewness**, several measures of which are available. Many models assume a Normal distribution, where, among other things, the data are symmetric around the mean.

Skewness measures asymmetry in the distribution, where left skew ($\text{mean} < \text{median}$) is indicated by negative skewness values, while right skew ($\text{mean} > \text{median}$) is indicated by positive values. The skew value will be near zero for data that follow a symmetric distribution.

9.13.1 Non-parametric Skewness

A simpler measure of skew, sometimes called the **nonparametric skew** and closely related to Pearson's notion of median skewness, falls between -1 and +1 for any distribution. It is just the difference between the mean and the median, divided by the standard deviation.

- Values greater than +0.2 are sometimes taken to indicate fairly substantial right skew, while values below -0.2 indicate fairly substantial left skew.

```
(mean(nnyfs$energy) - median(nnyfs$energy))/sd(nnyfs$energy)
```

```
[1] 0.114427
```

The [Wikipedia page on skewness](#), from which some of this material is derived, provides definitions for several other skewness measures.

9.14 Assessing Kurtosis (Heavy-Tailedness)

Another measure of a distribution's shape that can be found in the `psych` library is the **kurtosis**. Kurtosis is an indicator of whether the distribution is heavy-tailed or light-tailed as compared to a Normal distribution. Positive kurtosis means more of the variance is due to outliers - unusual points far away from the mean relative to what we might expect from a Normally distributed data set with the same standard deviation.

- A Normal distribution will have a kurtosis value near 0, a distribution with similar tail behavior to what we would expect from a Normal is said to be *mesokurtic*
- Higher kurtosis values (meaningfully higher than 0) indicate that, as compared to a Normal distribution, the observed variance is more the result of extreme outliers (i.e. heavy tails) as opposed to being the result of more modest sized deviations from the mean. These heavy-tailed, or outlier prone, distributions are sometimes called *leptokurtic*.
- Kurtosis values meaningfully lower than 0 indicate light-tailed data, with fewer outliers than we'd expect in a Normal distribution. Such distributions are sometimes referred to as *platykurtic*, and include distributions without outliers, like the Uniform distribution.

Here's a table:

Fewer outliers than a Normal	Approximately Normal	More outliers than a Normal
Light-tailed <i>platykurtic</i> (kurtosis < 0)	“Normalish” <i>mesokurtic</i> (kurtosis = 0)	Heavy-tailed <i>leptokurtic</i> (kurtosis > 0)

```
psych::kurtosi(nnyfs$energy)
```

```
[1] 1.130539
```

Note that the `kurtosi()` function is strangely named, and is part of the `psych` package.

9.14.1 The Standard Error of the Sample Mean

The **standard error** of the sample mean, which is the standard deviation divided by the square root of the sample size:

```
sd(nnyfs$energy)/sqrt(length(nnyfs$energy))
```

```
[1] 18.54018
```

9.15 The `describe` function in the `Hmisc` package

The `Hmisc` package has lots of useful functions. It's named for its main developer, Frank Harrell. The `describe` function in `Hmisc` knows enough to separate numerical from categorical variables, and give you separate (and detailed) summaries for each.

- For a categorical variable, it provides counts of total observations (n), the number of missing values, and the number of unique categories, along with counts and percentages falling in each category.
- For a numerical variable, it provides:
 - counts of total observations (n), the number of missing values, and the number of unique values
 - an Info value for the data, which indicates how continuous the variable is (a score of 1 is generally indicative of a completely continuous variable with no ties, while scores near 0 indicate lots of ties, and very few unique values)
 - the sample Mean

- Gini's mean difference, which is a robust measure of spread, with larger values indicating greater dispersion in the data. It is defined as the mean absolute difference between any pairs of observations.
- many sample percentiles (quantiles) of the data, specifically (5, 10, 25, 50, 75, 90, 95, 99)
- either a complete table of all observed values, with counts and percentages (if there are a modest number of unique values), or
- a table of the five smallest and five largest values in the data set, which is useful for range checking

```
nnyfs |>
  select(waist, energy, bmi) |>
  Hmisc::describe()
```

```
select(nnyfs, waist, energy, bmi)
```

```
3 Variables      1518 Observations
```

waist

	n	missing	distinct	Info	Mean	Gmd	.05	.10
1512		6	510	1	67.71	16.6	49.40	51.40
.25		.50	.75	.90	.95			
55.60		64.80	76.60	88.70	96.84			

```
lowest : 42.5 43.4 44.1 44.4 44.5, highest: 125.8 126.0 127.0 132.3 144.7
```

energy

	n	missing	distinct	Info	Mean	Gmd	.05	.10
1518		0	1137	1	1877	796.1	849	1047
.25		.50	.75	.90	.95			
1368		1794	2306	2795	3195			

```
lowest : 257 260 326 349 392, highest: 4382 4529 5085 5215 5265
```

bmi

	n	missing	distinct	Info	Mean	Gmd	.05	.10
1514		4	225	1	19.63	5.269	14.30	14.90
.25		.50	.75	.90	.95			
15.90		18.10	21.90	26.27	30.20			

```
lowest : 11.9 12.6 12.7 12.9 13.0, highest: 42.8 43.0 46.9 48.2 48.3
```

More on the `Info` value in `Hmisc::describe` is [available here](#)

9.16 Summarizing data within subgroups

Suppose we want to understand how the subjects whose diet involved consuming much more than usual yesterday compare to those who consumer their usual amount, or to those who consumed much less than usual, in terms of the energy they consumed, as well as the protein. We might start by looking at the medians and means.

```
nnyfs |>
  group_by(diet_yesterday) |>
  select(diet_yesterday, energy, protein) |>
  summarise_all(list(median = median, mean = mean))

# A tibble: 4 x 5
  diet_yesterday      energy_median protein_median energy_mean protein_mean
  <fct>                <dbl>        <dbl>       <dbl>        <dbl>
1 1_Much more than usual     2098       69.4      2150.       75.1
2 2_Usual                  1794       61.3      1858.       67.0
3 3_Much less than usual    1643       53.9      1779.       60.1
4 <NA>                     4348       155.      4348        155.
```

Perhaps we should restrict ourselves to the people who were not missing the `diet_yesterday` category, and look now at their `sugar` and `water` consumption.

```
nnyfs |>
  filter(complete.cases(diet_yesterday)) |>
  group_by(diet_yesterday) |>
  select(diet_yesterday, energy, protein, sugar, water) |>
  summarise_all(list(median))

# A tibble: 3 x 5
  diet_yesterday      energy protein sugar water
  <fct>                <dbl>   <dbl> <dbl> <dbl>
1 1_Much more than usual     2098    69.4  137.  500
2 2_Usual                  1794    61.3  114.  385.
3 3_Much less than usual    1643    53.9  115.  311.
```

It looks like the children in the “Much more than usual” category consumed more energy, protein, sugar and water than the children in the other two categories. Let’s draw a picture of this.

```
temp_dat <- nnyfs |>
  filter(complete.cases(diet_yesterday)) |>
  mutate(diet_yesterday = fct_recode(diet_yesterday,
    "Much more" = "1_Much more than usual",
    "Usual diet" = "2_Usual",
    "Much less" = "3_Much less than usual"))

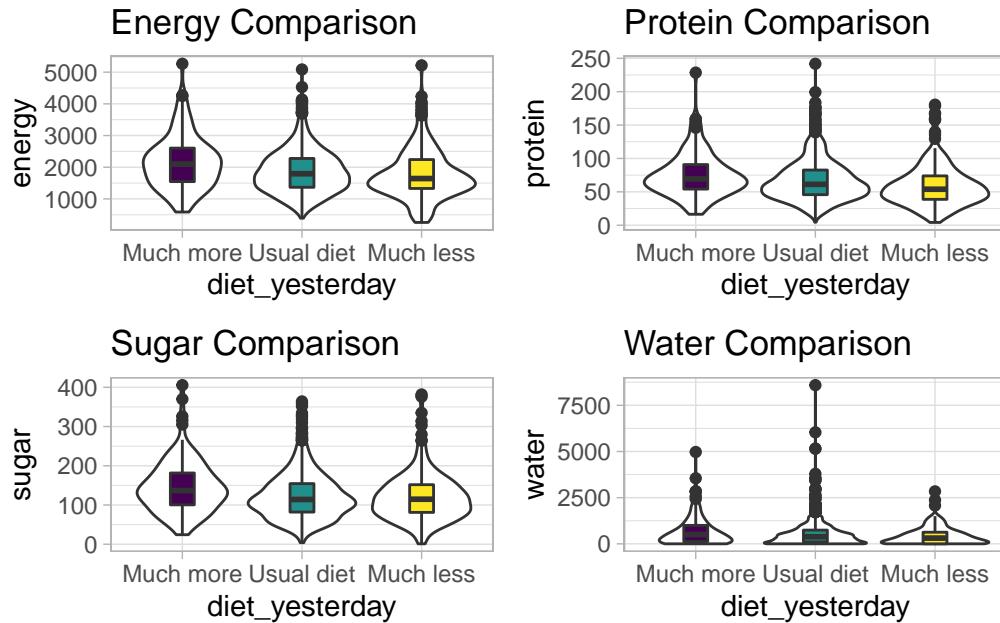
p1 <- ggplot(temp_dat, aes(x = diet_yesterday, y = energy)) +
  geom_violin() +
  geom_boxplot(aes(fill = diet_yesterday), width = 0.2) +
  theme_light() +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  labs(title = "Energy Comparison")

p2 <- ggplot(temp_dat, aes(x = diet_yesterday, y = protein)) +
  geom_violin() +
  geom_boxplot(aes(fill = diet_yesterday), width = 0.2) +
  theme_light() +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  labs(title = "Protein Comparison")

p3 <- ggplot(temp_dat, aes(x = diet_yesterday, y = sugar)) +
  geom_violin() +
  geom_boxplot(aes(fill = diet_yesterday), width = 0.2) +
  theme_light() +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  labs(title = "Sugar Comparison")

p4 <- ggplot(temp_dat, aes(x = diet_yesterday, y = water)) +
  geom_violin() +
  geom_boxplot(aes(fill = diet_yesterday), width = 0.2) +
  theme_light() +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  labs(title = "Water Comparison")
```

p1 + p2 + p3 + p4



We can see that there is considerable overlap in these distributions, regardless of what we're measuring.

9.17 Another Example

Suppose now that we ask a different question. Do kids in larger categories of BMI have larger waist circumferences?

```
nyfs |>
  group_by(bmi_cat) |>
  summarise(mean = mean(waist), sd = sd(waist),
            median = median(waist),
            skew_1 = round((mean(waist) - median(waist)) /
                           sd(waist), 2))

# A tibble: 5 x 5
  bmi_cat      mean     sd median skew_1
  <fct>       <dbl>   <dbl>  <dbl>    <dbl>
```

```

<fct>      <dbl> <dbl> <dbl> <dbl>
1 1_Underweight 55.2 7.58 54.5 0.09
2 2_Normal      NA   NA   NA   NA
3 3_Overweight  72.3 11.9 74   -0.14
4 4_Obese        NA   NA   NA   NA
5 <NA>          NA   NA   NA   NA

```

Oops. Looks like we need to filter for cases with complete data on both BMI category and waist circumference in order to get meaningful results. We should add a count, too.

```

nnyfs |>
  filter(complete.cases(bmi_cat, waist)) |>
  group_by(bmi_cat) |>
  summarise(count = n(), mean = mean(waist),
            sd = sd(waist), median = median(waist),
            skew_1 =
              round((mean(waist) - median(waist)) / sd(waist), 2))

# A tibble: 4 x 6
  bmi_cat     count    mean     sd median skew_1
<fct>      <int> <dbl> <dbl> <dbl> <dbl>
1 1_Underweight 41    55.2  7.58  54.5  0.09
2 2_Normal      917   61.2  9.35  59.5  0.19
3 3_Overweight  258   72.3  11.9   74   -0.14
4 4_Obese        294   85.6  17.1   86.8 -0.07

```

Or, we could use something like `favstats` from the `mosaic` package, which automatically accounts for missing data, and omits it when calculating summary statistics within each group.

```

mosaic::favstats(waist ~ bmi_cat, data = nnyfs) |>
  kable(digits = 1)

```

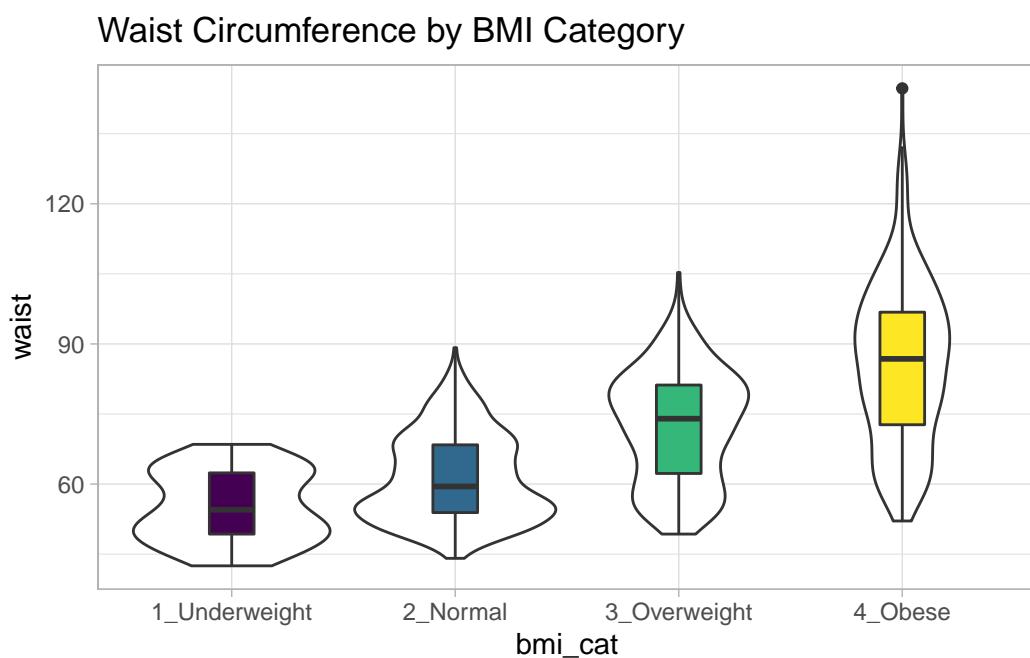
bmi_cat	min	Q1	median	Q3	max	mean	sd	n	missing
1_Underweight	42.5	49.3	54.5	62.4	68.5	55.2	7.6	41	0
2_Normal	44.1	53.9	59.5	68.4	89.2	61.2	9.4	917	3
3_Overweight	49.3	62.3	74.0	81.2	105.3	72.3	11.9	258	0
4_Obese	52.1	72.7	86.8	96.8	144.7	85.6	17.1	294	1

While patients in the heavier groups generally had higher waist circumferences, the standard deviations suggest there may be some meaningful overlap. Let's draw the picture, in this case a comparison boxplot accompanying a violin plot.

```

nnyfs |>
  filter(complete.cases(bmi_cat, waist)) %>%
  ggplot(., aes(x = bmi_cat, y = waist)) +
  geom_violin() +
  geom_boxplot(aes(fill = bmi_cat), width = 0.2) +
  theme_light() +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  labs(title = "Waist Circumference by BMI Category")

```



Note the use of the `%>%` pipe here. I need to write more about this.

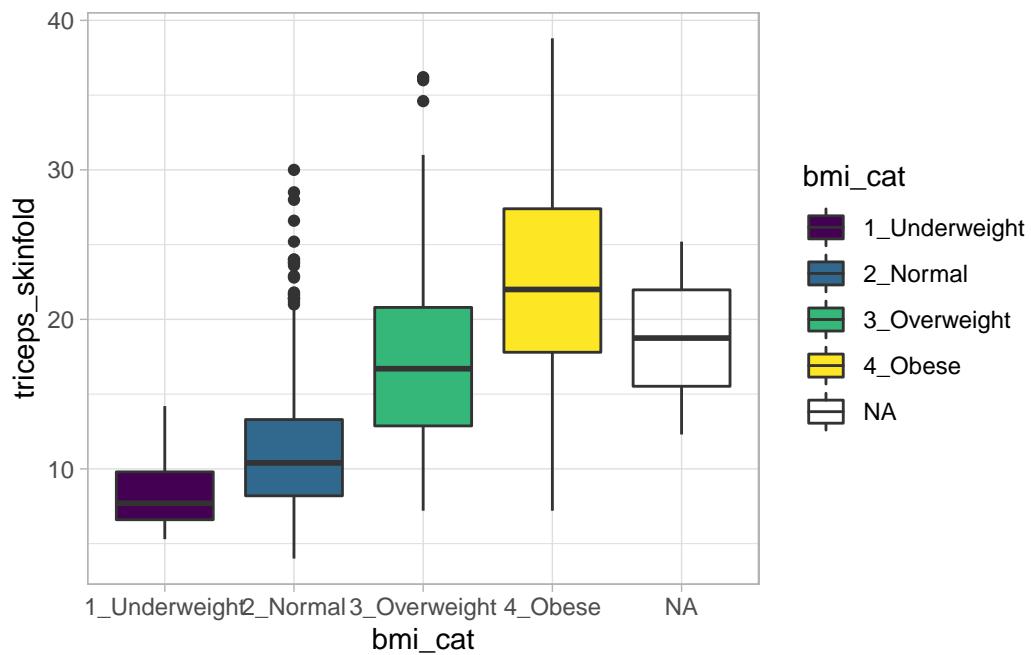
The data transformation with dplyr cheat sheet found under the Help menu in RStudio is a great resource. And, of course, for more details, visit Wickham and Grolemund (2022).

9.18 Boxplots to Relate an Outcome to a Categorical Predictor

Boxplots are much more useful when comparing samples of data. For instance, consider this comparison boxplot describing the triceps skinfold results across the four levels of BMI category.

```
ggplot(nnyfs, aes(x = bmi_cat, y = triceps_skinfold,
                  fill = bmi_cat)) +
  geom_boxplot() +
  scale_fill_viridis_d() +
  theme_light()
```

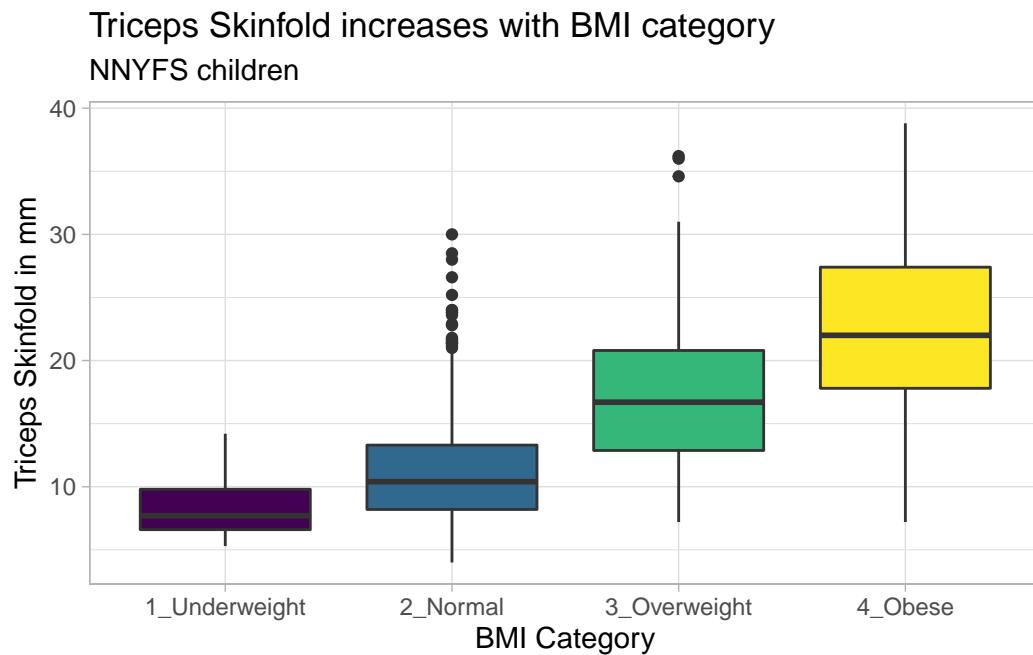
Warning: Removed 21 rows containing non-finite values (stat_boxplot).



Again, we probably want to omit those missing values (both in `bmi_cat` and `triceps_skinfold`) and also eliminate the repetitive legend (guides) on the right.

```
nnyfs |>
  filter(complete.cases(bmi_cat, triceps_skinfold)) %>%
  ggplot(., aes(x = bmi_cat, y = triceps_skinfold,
                 fill = bmi_cat)) +
  geom_boxplot() +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  theme_light() +
  labs(x = "BMI Category", y = "Triceps Skinfold in mm",
```

```
title = "Triceps Skinfold increases with BMI category",
subtitle = "NNYFS children")
```



As always, the boxplot shows the five-number summary (minimum, 25th percentile, median, 75th percentile and maximum) in addition to highlighting candidate outliers.

9.18.1 Augmenting the Boxplot with the Sample Mean

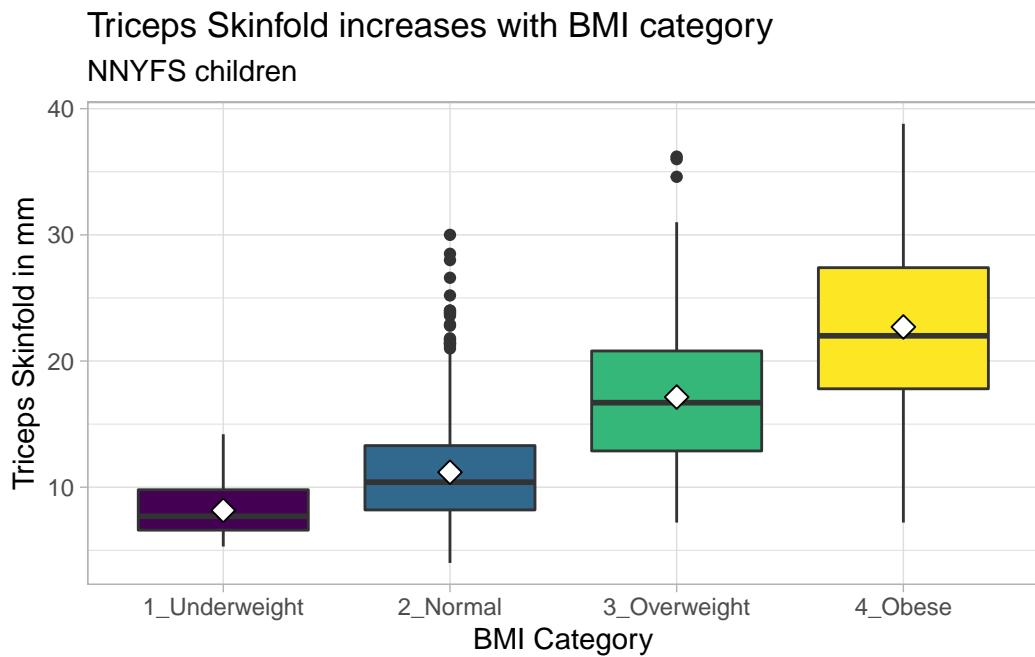
Often, we want to augment such a plot, perhaps by adding a little diamond to show the **sample mean** within each category, so as to highlight skew (in terms of whether the mean is meaningfully different from the median.)

```
nnyfs |>
  filter(complete.cases(bmi_cat, triceps_skinfold)) %>%
  ggplot(., aes(x = bmi_cat, y = triceps_skinfold,
                fill = bmi_cat)) +
  geom_boxplot() +
  stat_summary(fun="mean", geom="point",
              shape=23, size=3, fill="white") +
  scale_fill_viridis_d()
```

```

guides(fill = "none") +
theme_light() +
labs(x = "BMI Category", y = "Triceps Skinfold in mm",
title = "Triceps Skinfold increases with BMI category",
subtitle = "NNYFS children")

```



9.19 Building a Violin Plot

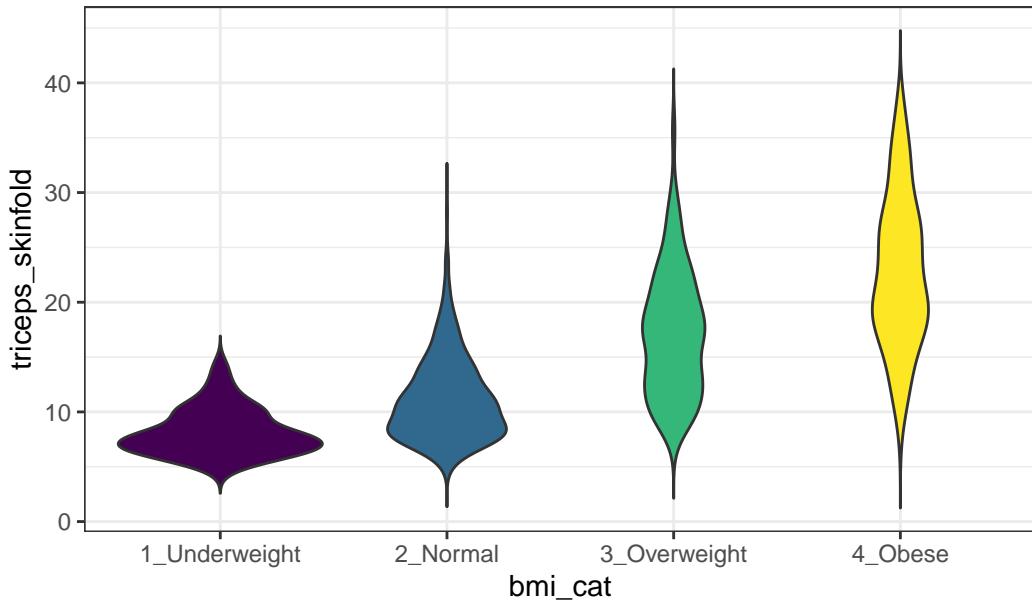
There are a number of other plots which compare distributions of data sets. An interesting one is called a **violin plot**. A violin plot is a kernel density estimate, mirrored to form a symmetrical shape.

```

nnyfs |>
filter(complete.cases(triceps_skinfold, bmi_cat)) %>%
ggplot(., aes(x=bmi_cat, y=triceps_skinfold,
fill = bmi_cat)) +
geom_violin(trim=FALSE) +
scale_fill_viridis_d() +
guides(fill = "none") +
labs(title = "Triceps Skinfold by BMI Category")

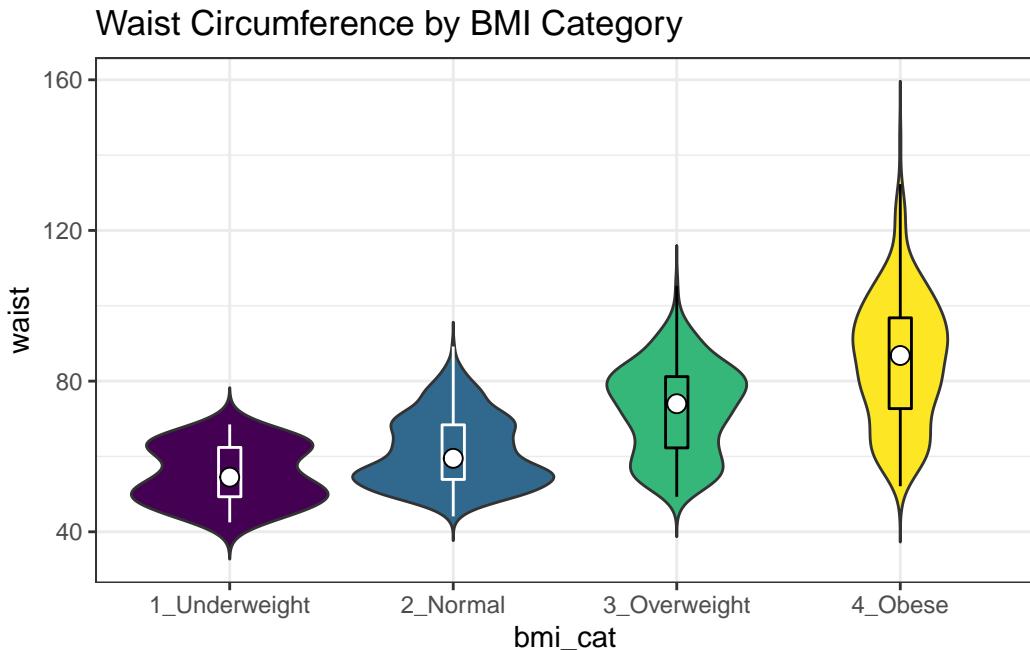
```

Triceps Skinfold by BMI Category



Traditionally, these plots are shown with overlaid boxplots and a white dot at the median, like this example, now looking at waist circumference again.

```
nnyfs |>
  filter(complete.cases(waist, bmi_cat)) %>%
  ggplot(., aes(x = bmi_cat, y = waist,
                 fill = bmi_cat)) +
  geom_violin(trim=FALSE) +
  geom_boxplot(width=.1, outlier.colour=NA,
               color = c(rep("white",2), rep("black",2))) +
  stat_summary(fun=median, geom="point",
               fill="white", shape=21, size=3) +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  labs(title = "Waist Circumference by BMI Category")
```

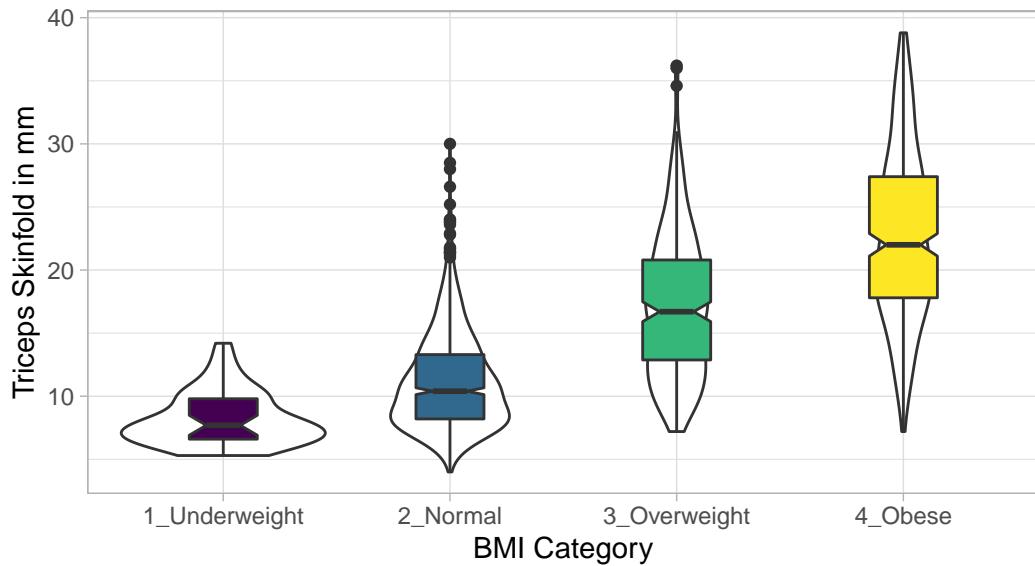


9.19.1 Adding Notches to a Boxplot

Notches are used in boxplots to help visually assess whether the medians of the distributions across the various groups actually differ to a statistically detectable extent. Think of them as confidence regions around the medians. If the notches do not overlap, as in this situation, this provides some evidence that the medians in the populations represented by these samples may be different.

```
nnyfs |>
  filter(complete.cases(bmi_cat, triceps_skinfold)) %>%
  ggplot(., aes(x = bmi_cat, y = triceps_skinfold)) +
  geom_violin() +
  geom_boxplot(aes(fill = bmi_cat), width = 0.3, notch = TRUE) +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  theme_light() +
  labs(x = "BMI Category", y = "Triceps Skinfold in mm",
       title = "Triceps Skinfold increases with BMI category",
       subtitle = "NNYFS children")
```

Triceps Skinfold increases with BMI category
NNYFS children

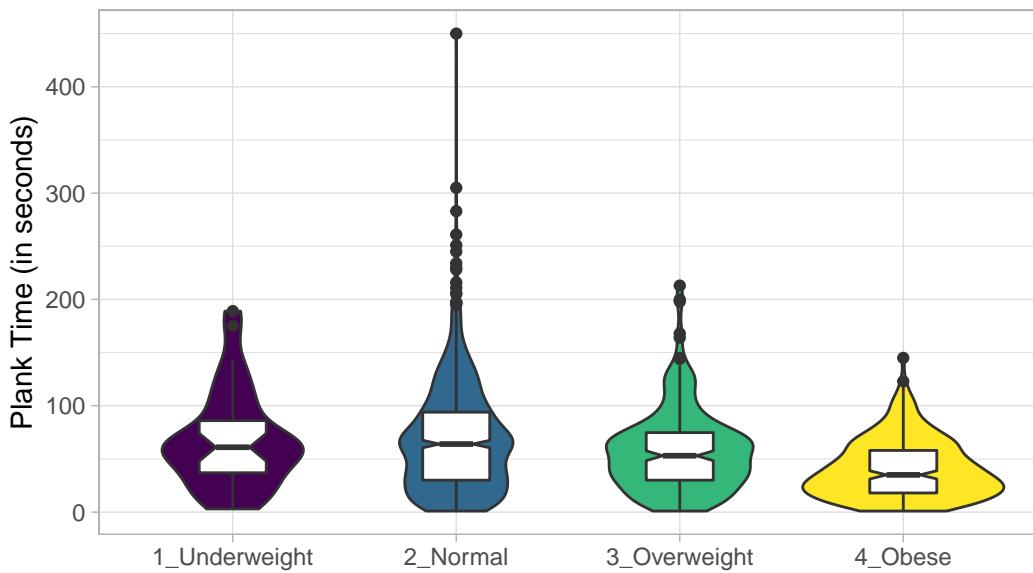


There is no overlap between the notches for each of the four categories, so we might reasonably conclude that the true median triceps skinfold values across the four categories are statistically significantly different.

For an example where the notches do overlap, consider the comparison of plank times by BMI category.

```
nnyfs |>
  filter(complete.cases(bmi_cat, plank_time)) %>%
  ggplot(., aes(x=bmi_cat, y=plank_time)) +
  geom_violin(aes(fill = bmi_cat)) +
  geom_boxplot(width = 0.3, notch=TRUE) +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  theme_light() +
  labs(title = "Plank Times by BMI category",
       x = "", y = "Plank Time (in seconds)")
```

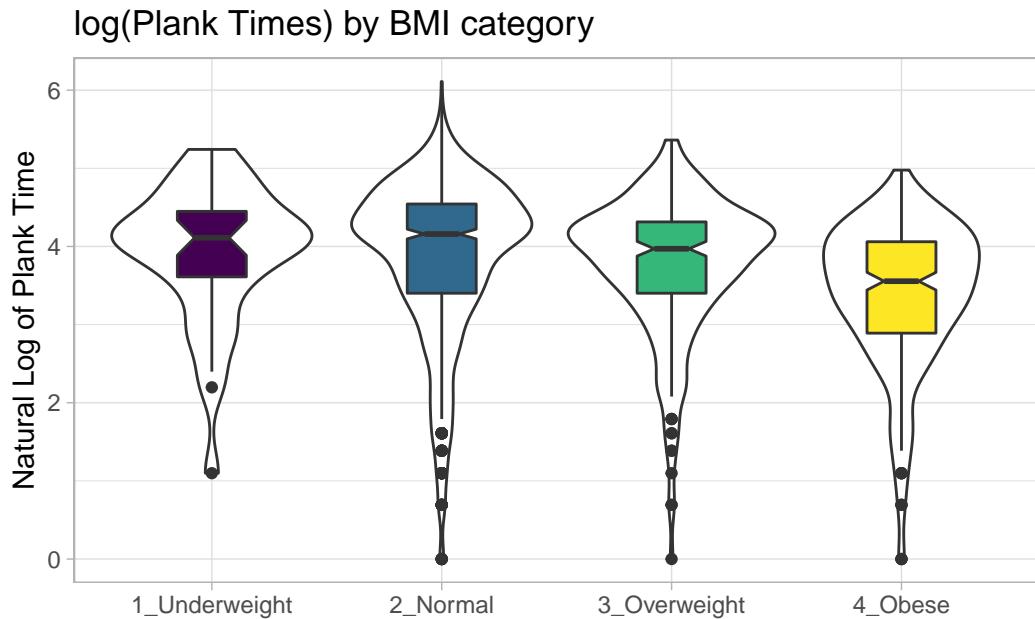
Plank Times by BMI category



The overlap in the notches (for instance between Underweight and Normal) suggests that the median plank times in the population of interest don't necessarily differ in a meaningful way by BMI category, other than perhaps the Obese group which may have a shorter time.

These data are somewhat right skewed. Would a logarithmic transformation in the plot help us see the patterns more clearly?

```
nyfs |>
  filter(complete.cases(bmi_cat, plank_time)) %>%
  ggplot(., aes(x=bmi_cat, y = log(plank_time))) +
  geom_violin() +
  geom_boxplot(aes(fill = bmi_cat), width = 0.3, notch=TRUE) +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  theme_light() +
  labs(title = "log(Plank Times) by BMI category",
       x = "", y = "Natural Log of Plank Time")
```

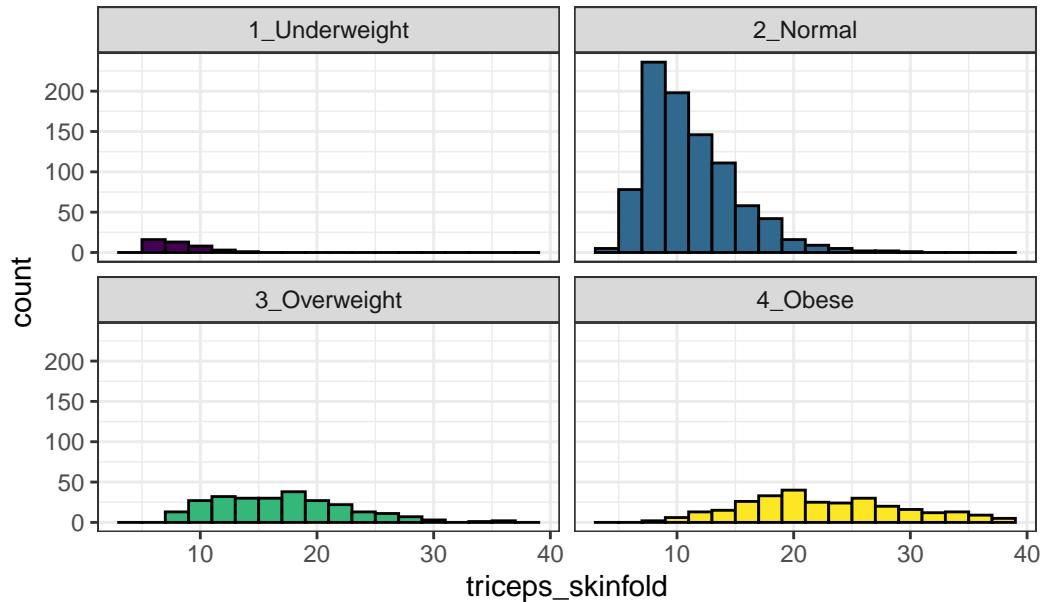


9.20 Using Multiple Histograms to Make Comparisons

We can make an array of histograms to describe multiple groups of data, using `ggplot2` and the notion of **faceting** our plot.

```
nnysf |>
  filter(complete.cases(triceps_skinfold, bmi_cat)) %>%
  ggplot(., aes(x=triceps_skinfold, fill = bmi_cat)) +
  geom_histogram(binwidth = 2, color = "black") +
  facet_wrap(~ bmi_cat) +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  labs(title = "Triceps Skinfold by BMI Category")
```

Triceps Skinfold by BMI Category

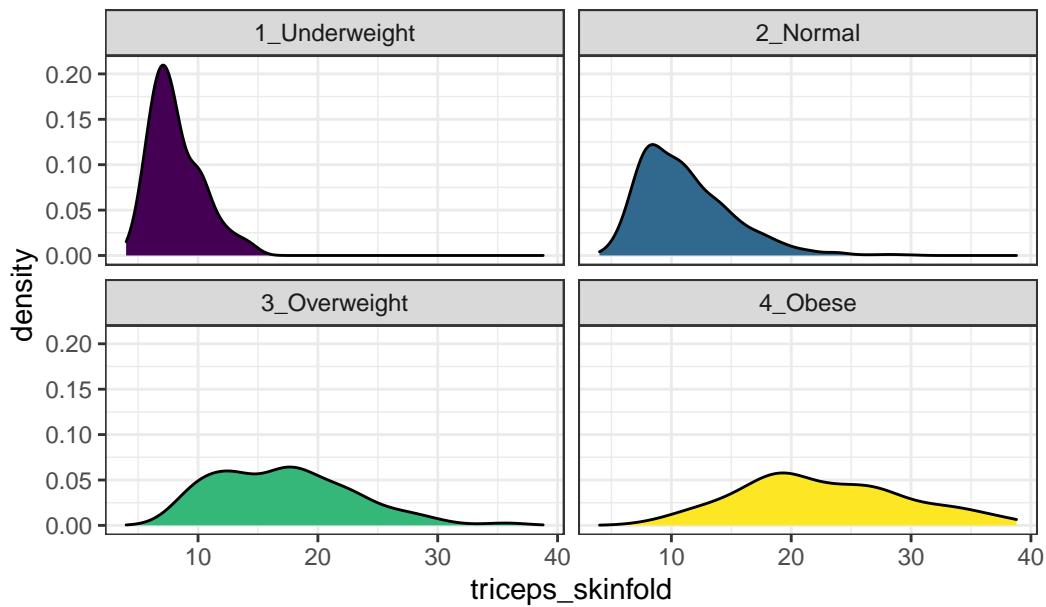


9.21 Using Multiple Density Plots to Make Comparisons

Or, we can make a series of density plots to describe multiple groups of data.

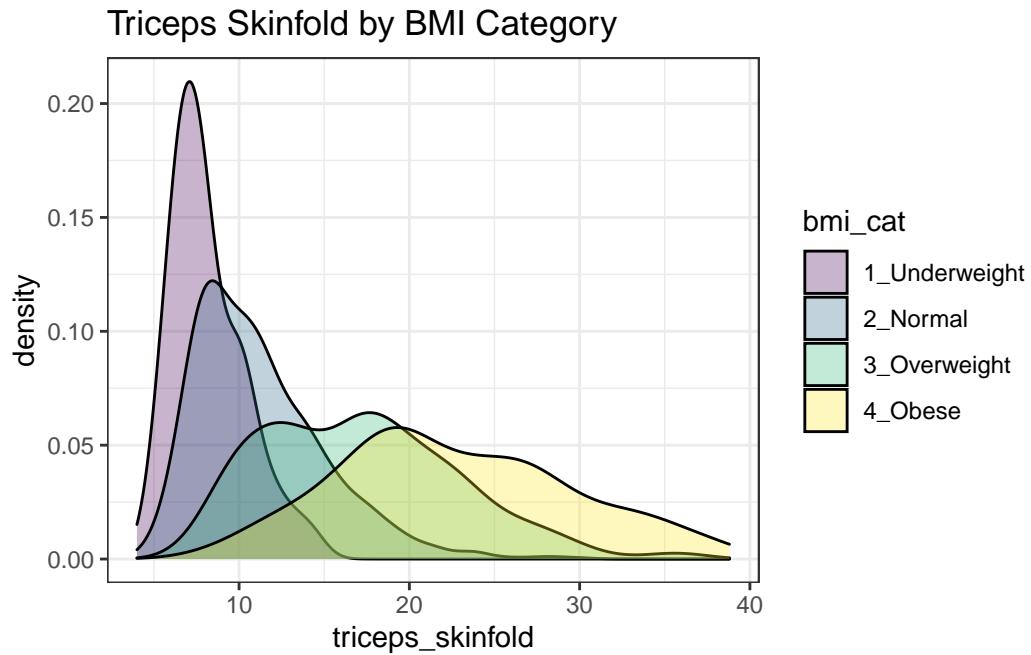
```
nyfs |>
  filter(complete.cases(triceps_skinfold, bmi_cat)) %>%
  ggplot(., aes(x=triceps_skinfold, fill = bmi_cat)) +
  geom_density(color = "black") +
  facet_wrap(~ bmi_cat) +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  labs(title = "Triceps Skinfold by BMI Category")
```

Triceps Skinfold by BMI Category



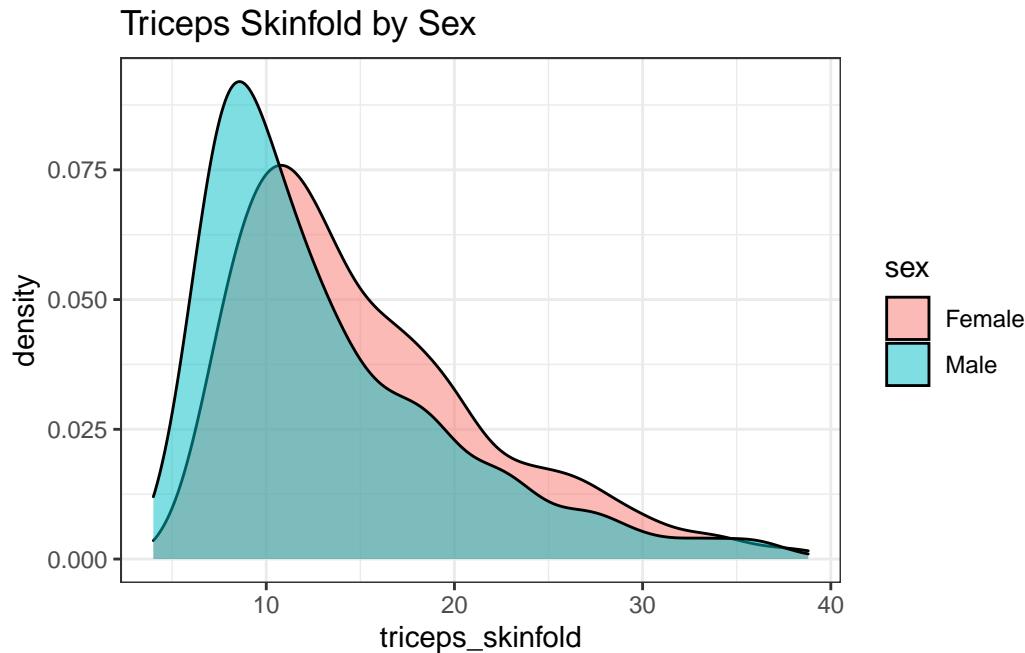
Or, we can plot all of the densities on top of each other with semi-transparent fills.

```
nyfs |>
  filter(complete.cases(triceps_skinfold, bmi_cat)) %>%
  ggplot(., aes(x=triceps_skinfold, fill = bmi_cat)) +
  geom_density(alpha=0.3) +
  scale_fill_viridis_d() +
  labs(title = "Triceps Skinfold by BMI Category")
```



This really works better when we are comparing only two groups, like females to males.

```
nnyfs |>
  filter(complete.cases(triceps_skinfold, sex)) %>%
  ggplot(., aes(x=triceps_skinfold, fill = sex)) +
  geom_density(alpha=0.5) +
  labs(title = "Triceps Skinfold by Sex")
```



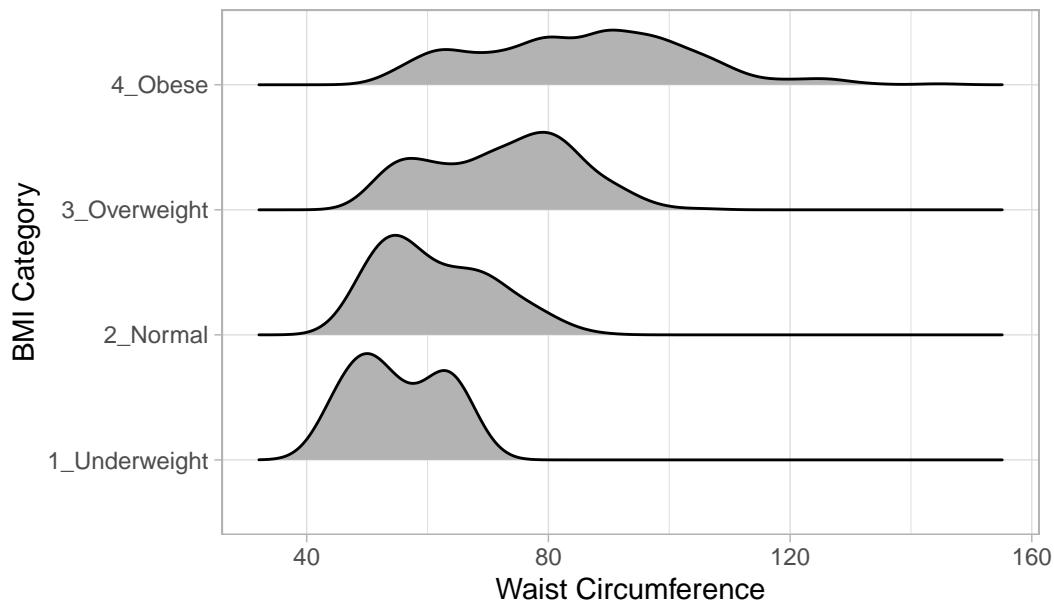
9.22 A Ridgeline Plot

Some people don't like violin plots - for example, see <https://simplystatistics.org/2017/07/13/the-joy-of-no-more-violin-plots/>. An alternative plot is available as part of the `ggridges` package. This shows the distribution of several groups simultaneously, especially when you have lots of subgroup categories, and is called a **ridgeline plot**.

```
nnyfs |>
  filter(complete.cases(waist, bmi_cat)) %>%
  ggplot(., aes(x = waist, y = bmi_cat, height = ..density..)) +
  ggridges::geom_density_ridges(scale = 0.85) +
  theme_light() +
  labs(title = "Ridgeline Plot of Waist Circumference by BMI category (nnyfs)",
       x = "Waist Circumference", y = "BMI Category")
```

Picking joint bandwidth of 3.47

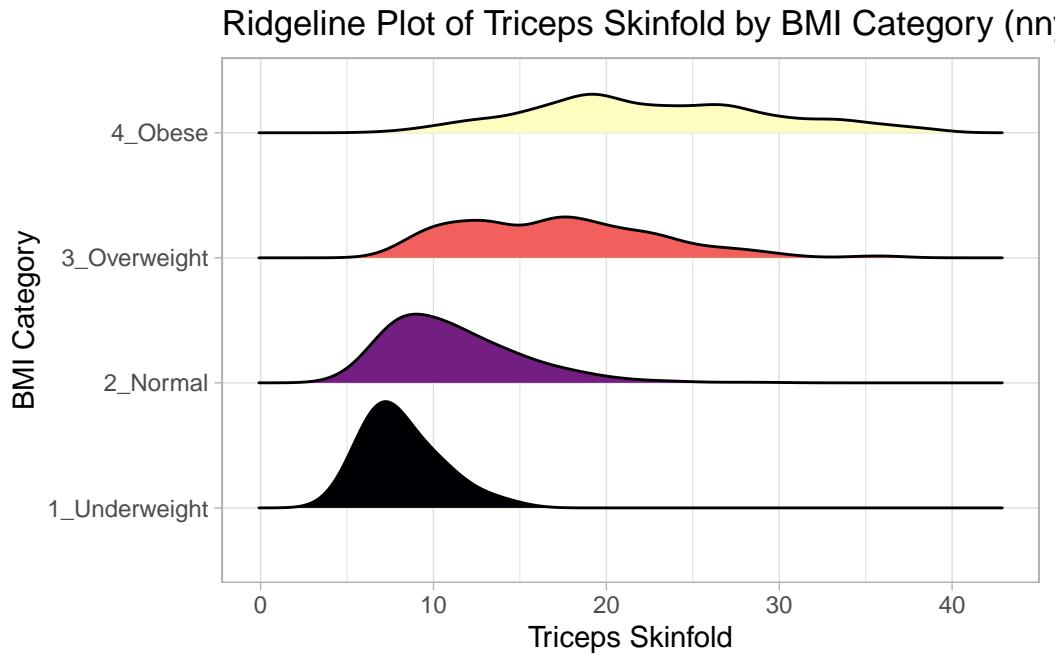
Ridgeline Plot of Waist Circumference by BMI category



And here's a ridgeline plot for the triceps skinfolds. We'll start by sorting the subgroups by the median value of our outcome (triceps skinfold) in this case, though it turns out not to matter. We'll also add some color.

```
nnyfs |>
  filter(complete.cases(bmi_cat, triceps_skinfold)) |>
  mutate(bmi_cat = fct_reorder(bmi_cat,
                                triceps_skinfold,
                                .fun = median)) %>%
  ggplot(., aes(x = triceps_skinfold, y = bmi_cat,
                fill = bmi_cat, height = ..density..)) +
  ggridges::geom_density_ridges(scale = 0.85) +
  scale_fill_viridis_d(option = "magma") +
  guides(fill = "none") +
  labs(title = "Ridgeline Plot of Triceps Skinfold by BMI Category (nnyfs)",
       x = "Triceps Skinfold", y = "BMI Category") +
  theme_light()
```

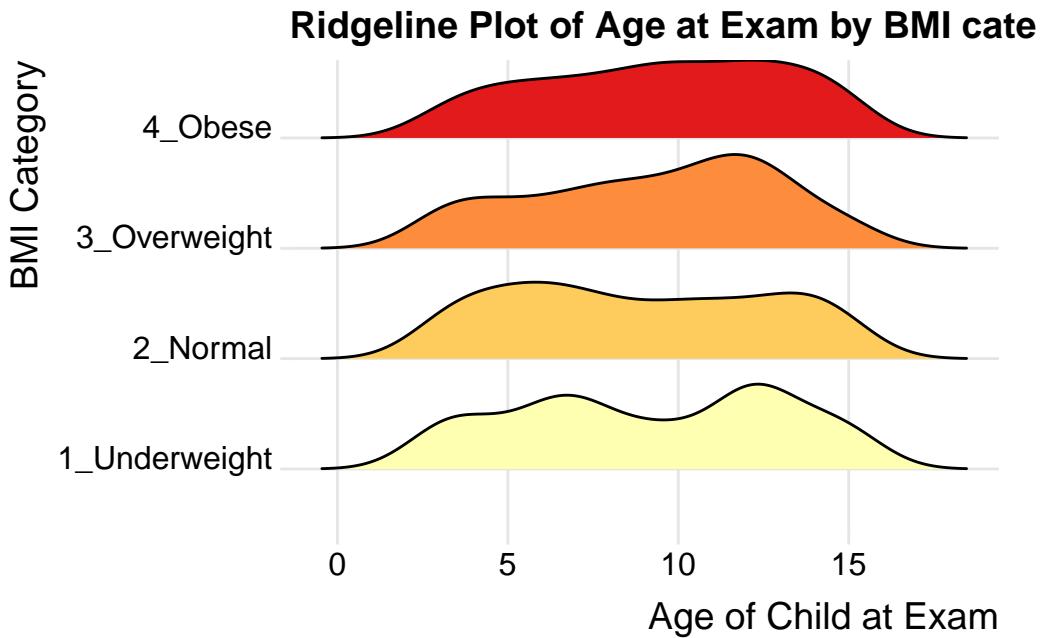
Picking joint bandwidth of 1.37



For one last example, we'll look at age by BMI category, so that sorting the BMI subgroups by the median matters, and we'll try an alternate color scheme, and a theme specially designed for the ridgeline plot.

```
nnyfs |>
  filter(complete.cases(bmi_cat, age_child)) |>
  mutate(bmi_cat = reorder(bmi_cat, age_child, median)) %>%
  ggplot(aes(x = age_child, y = bmi_cat, fill = bmi_cat, height = ..density..)) +
  ggridges::geom_density_ridges(scale = 0.85) +
  scale_fill_brewer(palette = "YlOrRd") +
  guides(fill = "none") +
  labs(title = "Ridgeline Plot of Age at Exam by BMI category (nnyfs)",
       x = "Age of Child at Exam", y = "BMI Category") +
  ggridges::theme_ridges()
```

Picking joint bandwidth of 1.15



9.23 What Summaries to Report

It is usually helpful to focus on the shape, center and spread of a distribution. Bock, Velleman and DeVeaux provide some useful advice:

- If the data are skewed, report the median and IQR (or the three middle quantiles). You may want to include the mean and standard deviation, but you should point out why the mean and median differ. The fact that the mean and median do not agree is a sign that the distribution may be skewed. A histogram will help you make that point.
- If the data are symmetric, report the mean and standard deviation, and possibly the median and IQR as well.
- If there are clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be revealing. The median and IQR are not likely to be seriously affected by outliers.

10 Assessing Normality

10.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

We also use the `favstat` function from the `mosaic` package in this chapter, but do not load the whole packages.

10.2 Introduction

Data are well approximated by a Normal distribution if the shape of the data's distribution is a good match for a Normal distribution with mean and standard deviation equal to the sample statistics.

- the data are symmetrically distributed about a single peak, located at the sample mean
- the spread of the distribution is well characterized by a Normal distribution with standard deviation equal to the sample standard deviation
- the data show outlying values (both in number of candidate outliers, and size of the distance between the outliers and the center of the distribution) that are similar to what would be predicted by a Normal model.

We have several tools for assessing Normality of a single batch of data, including:

- a histogram with superimposed Normal distribution
- histogram variants (like the boxplot) which provide information on the center, spread and shape of a distribution
- the Empirical Rule for interpretation of a standard deviation

- a specialized *normal Q-Q plot* (also called a normal probability plot or normal quantile-quantile plot) designed to reveal differences between a sample distribution and what we might expect from a normal distribution of a similar number of values with the same mean and standard deviation

10.3 Empirical Rule Interpretation of the Standard Deviation

For a set of measurements that follows a Normal distribution, the interval:

- Mean \pm Standard Deviation contains approximately 68% of the measurements;
- Mean \pm 2(Standard Deviation) contains approximately 95% of the measurements;
- Mean \pm 3(Standard Deviation) contains approximately all (99.7%) of the measurements.

Again, most data sets do not follow a Normal distribution. We will occasionally think about transforming or re-expressing our data to obtain results which are better approximated by a Normal distribution, in part so that a standard deviation can be more meaningful.

For the energy data we have been studying, here again are some summary statistics...

```
nnyfs <- read_rds("data/nnyfs.Rds")

mosaic::favstats(nnyfs$energy)

min      Q1 median      Q3    max      mean        sd      n missing
257  1367.5  1794.5  2306  5265  1877.157  722.3537  1518          0
```

The mean is 1877 and the standard deviation is 722, so if the data really were Normally distributed, we'd expect to see:

- About 68% of the data in the range (1155, 2600). In fact, 1085 of the 1518 energy values are in this range, or 71.5%.
- About 95% of the data in the range (432, 3322). In fact, 1450 of the 1518 energy values are in this range, or 95.5%.
- About 99.7% of the data in the range (-290, 4044). In fact, 1502 of the 1518 energy values are in this range, or 98.9%.

So, based on this Empirical Rule approximation, do the energy data seem to be well approximated by a Normal distribution?

10.4 Describing Outlying Values with Z Scores

The maximum energy consumption value here is 5265. One way to gauge how extreme this is (or how much of an outlier it is) uses that observation's **Z score**, the number of standard deviations away from the mean that the observation falls.

Here, the maximum value, 5265 is 4.69 standard deviations above the mean, and thus has a Z score of 4.7.

A negative Z score would indicate a point below the mean, while a positive Z score indicates, as we've seen, a point above the mean. The minimum body-mass index, 257 is 2.24 standard deviations *below* the mean, so it has a Z score of -2.2.

Recall that the Empirical Rule suggests that if a variable follows a Normal distribution, it would have approximately 95% of its observations falling inside a Z score of (-2, 2), and 99.74% falling inside a Z score range of (-3, 3).

10.4.1 Fences and Z Scores

Note the relationship between the fences (Tukey's approach to identifying points which fall within the whiskers of a boxplot, as compared to candidate outliers) and the Z scores.

The upper inner fence in this case falls at 3713.75, which indicates a Z score of 2.5, while the lower inner fence falls at -40.25, which indicates a Z score of -2.7. It is neither unusual nor inevitable for the inner fences to fall at Z scores near -2.0 and +2.0.

10.5 Comparing a Histogram to a Normal Distribution

Most of the time, when we want to understand whether our data are well approximated by a Normal distribution, we will use a graph to aid in the decision.

One option is to build a histogram with a Normal density function (with the same mean and standard deviation as our data) superimposed. This is one way to help visualize deviations between our data and what might be expected from a Normal distribution.

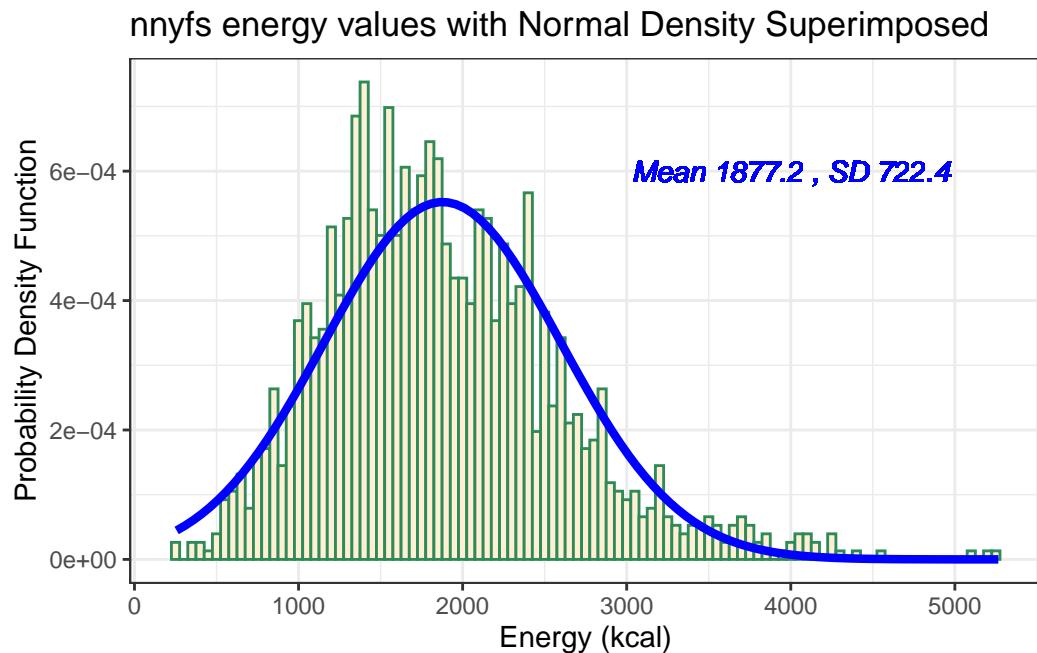
```
res <- mosaic::favstats(~ energy, data = nnyfs)
bin_w <- 50 # specify binwidth

ggplot(nnyfs, aes(x=energy)) +
  geom_histogram(aes(y = ..density..), binwidth = bin_w,
                 fill = "papayawhip", color = "seagreen") +
  stat_function(fun = dnorm,
```

```

    args = list(mean = res$mean, sd = res$sd),
    lwd = 1.5, col = "blue") +
  geom_text(aes(label = paste("Mean", round(res$mean,1),
                             ", SD", round(res$sd,1))),
            x = 4000, y = 0.0006,
            color="blue", fontface = "italic") +
  labs(title = "nnyfs energy values with Normal Density Superimposed",
       x = "Energy (kcal)", y = "Probability Density Function")

```



Does it seem as though the Normal model (as shown in the blue density curve) is an effective approximation to the observed distribution shown in the bars of the histogram?

We'll return shortly to the questions:

- Does a Normal distribution model fit our data well? *and*
- If the data aren't Normal, but we want to use a Normal model anyway, what should we do?

10.5.1 Histogram of energy with Normal model (with Counts)

But first, we'll demonstrate an approach to building a histogram of counts (rather than a probability density) and then superimposing a Normal model.

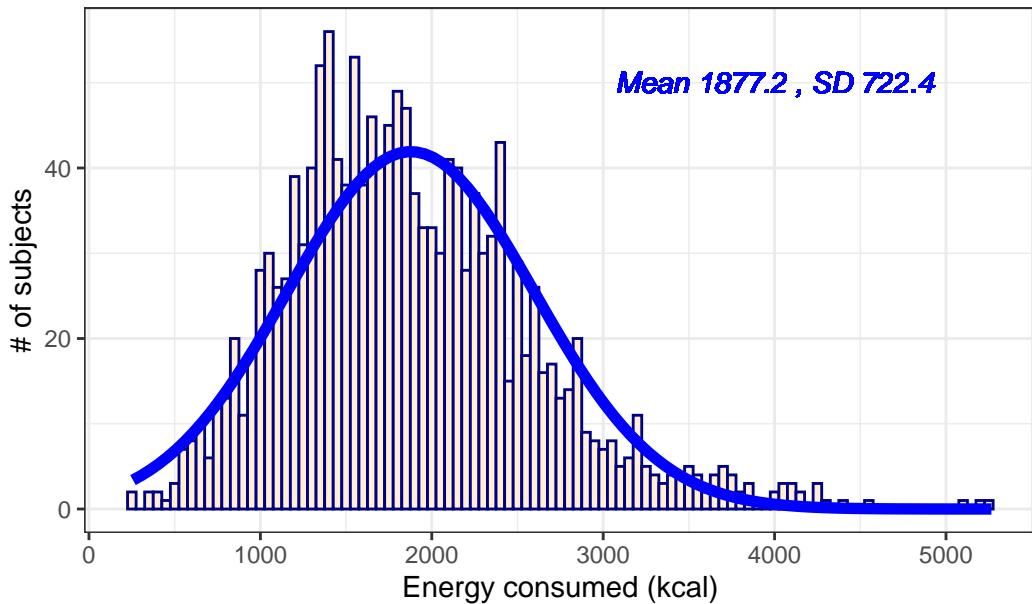
```

res <- mosaic::favstats(~ energy, data = nnyfs)
bin_w <- 50 # specify binwidth

ggplot(nnyfs, aes(x = energy)) +
  geom_histogram(binwidth = bin_w,
                 fill = "papayawhip",
                 col = "navy") +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) * res$n * bin_w,
    col = "blue", size = 2) +
  geom_text(aes(label = paste("Mean", round(res$mean,1),
                            ", SD", round(res$sd,1))),
            x = 4000, y = 50,
            color="blue", fontface = "italic") +
  labs(title = "Histogram of energy, with Normal Model",
       x = "Energy consumed (kcal)", y = "# of subjects")

```

Histogram of energy, with Normal Model



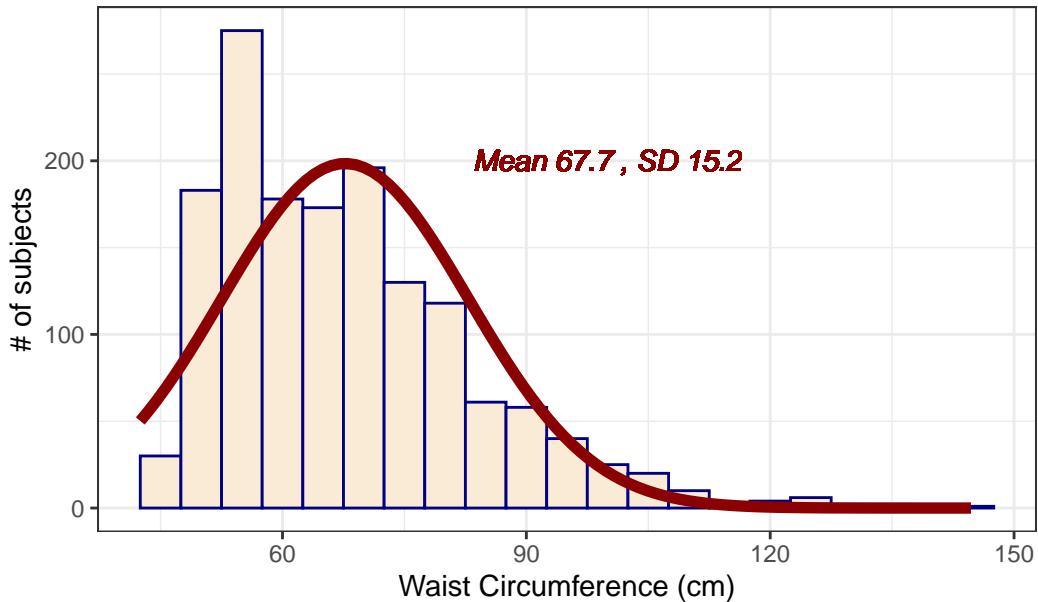
10.6 Does a Normal model work well for the waist circumference?

Now, suppose we instead look at the `waist` data, remembering to filter the data to the complete cases before plotting. Do these data appear to follow a Normal distribution?

```
res <- mosaic::favstats(~ waist, data = nnyfs)
bin_w <- 5 # specify binwidth

nnnyfs |> filter(complete.cases(waist)) %>%
  ggplot(., aes(x = waist)) +
  geom_histogram(binwidth = bin_w,
                 fill = "antiquewhite",
                 col = "navy") +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) *
      res$n * bin_w,
    col = "darkred", size = 2) +
  geom_text(aes(label = paste("Mean", round(res$mean,1),
                            ", SD", round(res$sd,1))),
            x = 100, y = 200,
            color="darkred", fontface = "italic") +
  labs(title = "Histogram of waist, with Normal Model",
       x = "Waist Circumference (cm)", y = "# of subjects")
```

Histogram of waist, with Normal Model



```
mosaic::favstats(~ waist, data = nnyfs)
```

min	Q1	median	Q3	max	mean	sd	n	missing
42.5	55.6	64.8	76.6	144.7	67.70536	15.19809	1512	6

The mean is 67.71 and the standard deviation is 15.2 so if the `waist` data really were Normally distributed, we'd expect to see:

- About 68% of the data in the range (52.51, 82.9). In fact, 1076 of the 1512 Age values are in this range, or 71.2%.
- About 95% of the data in the range (37.31, 98.1). In fact, 1443 of the 1512 Age values are in this range, or 95.4%.
- About 99.7% of the data in the range (22.11, 113.3). In fact, 1500 of the 1512 Age values are in this range, or 99.2%.

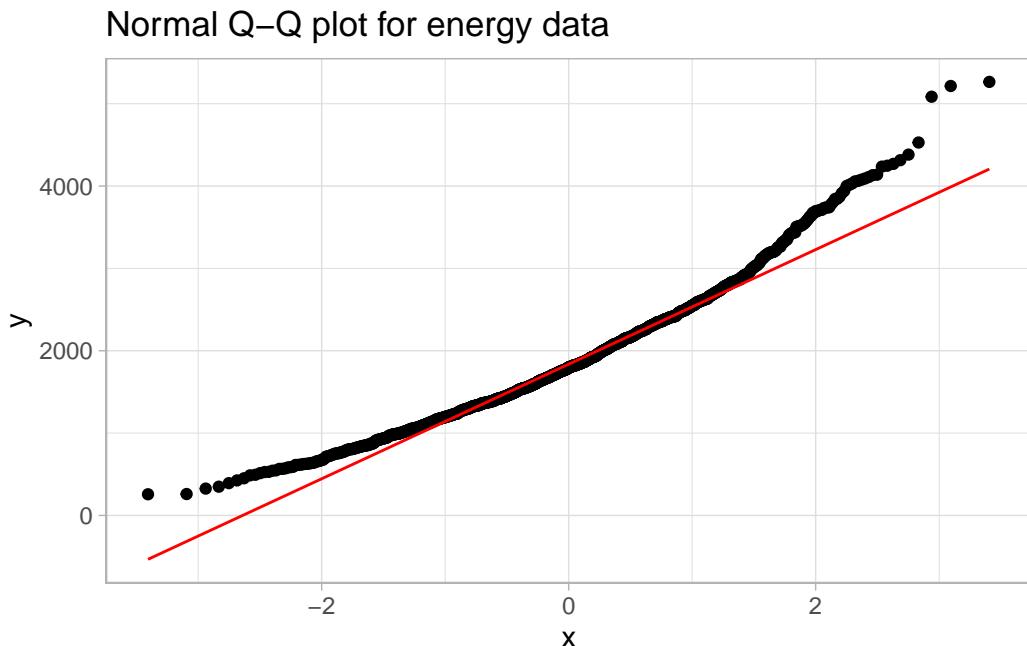
How does the Normal approximation work for waist circumference, according to the Empirical Rule?

10.7 The Normal Q-Q Plot

A normal probability plot (or normal quantile-quantile plot) of the energy results from the `nnyfs` data, developed using `ggplot2` is shown below. In this case, this is a picture of 1518 energy consumption assessments. The idea of a normal Q-Q plot is that it plots the observed sample values (on the vertical axis) and then, on the horizontal, the expected or theoretical quantiles that would be observed in a standard normal distribution (a Normal distribution with mean 0 and standard deviation 1) with the same number of observations.

A Normal Q-Q plot will follow a straight line when the data are (approximately) Normally distributed. When the data have a different shape, the plot will reflect that.

```
ggplot(nnyfs, aes(sample = energy)) +  
  geom_qq() + geom_qq_line(col = "red") +  
  theme_light() +  
  labs(title = "Normal Q-Q plot for energy data")
```



10.8 Interpreting the Normal Q-Q Plot

The purpose of a Normal Q-Q plot is to help point out distinctions from a Normal distribution. A Normal distribution is symmetric and has certain expectations regarding its tails. The

Normal Q-Q plot can help us identify data as well approximated by a Normal distribution, or not, because of:

- skew (including distinguishing between right skew and left skew)
- behavior in the tails (which could be heavy-tailed [more outliers than expected] or light-tailed)

10.8.1 Data from a Normal distribution shows up as a straight line in a Normal Q-Q plot

We'll demonstrate the looks that we can obtain from a Normal Q-Q plot in some simulations. First, here is an example of a Normal Q-Q plot, and its associated histogram, for a sample of 200 observations simulated from a Normal distribution.

```
set.seed(123431) # so the results can be replicated

# simulate 200 observations from a Normal(20, 5) distribution and place them
# in the d variable within the temp.1 data frame
temp.1 <- data.frame(d = rnorm(200, mean = 20, sd = 5))

# left plot - basic Normal Q-Q plot of simulated data
p1 <- ggplot(temp.1, aes(sample = d)) +
  geom_qq() + geom_qq_line(col = "red") +
  theme_light() +
  labs(y = "Ordered Simulated Sample Data")

# right plot - histogram with superimposed normal distribution
res <- mosaic::favstats(~ d, data = temp.1)
bin_w <- 2 # specify binwidth

p2 <- ggplot(temp.1, aes(x = d)) +
  geom_histogram(binwidth = bin_w,
                 fill = "papayawhip",
                 col = "seagreen") +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) *
      res$n * bin_w,
    col = "blue", size = 1.5) +
  geom_text(aes(label = paste("Mean", round(res$mean, 1),
```

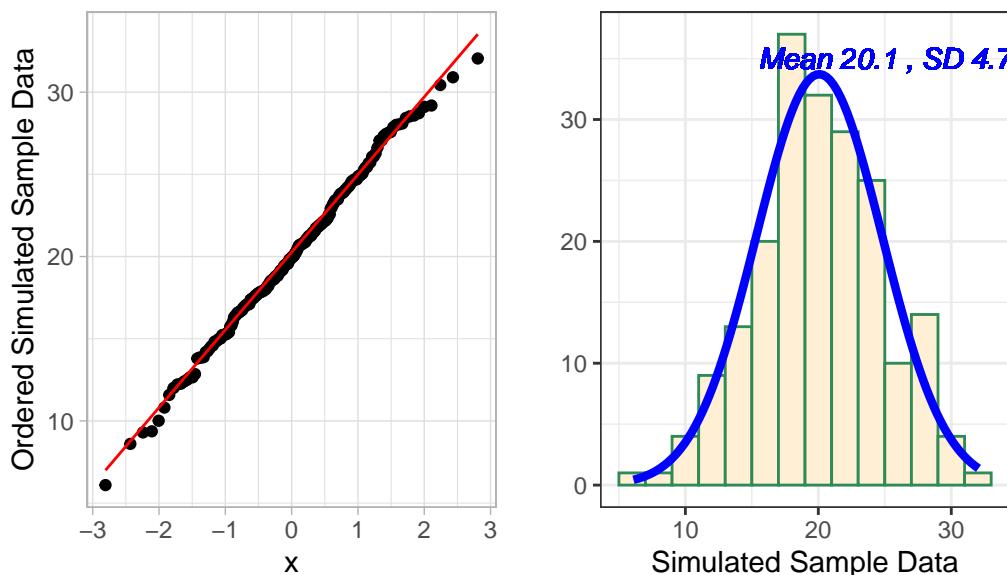
```

    ", SD", round(res$sd,1))),
x = 25, y = 35,
color="blue", fontface = "italic") +
labs(x = "Simulated Sample Data", y = "")

p1 + p2 +
plot_annotation(title = "200 observations from a simulated Normal distribution")

```

200 observations from a simulated Normal distribution



```
# uses patchwork package to combine plots
```

These simulated data appear to be well-modeled by the Normal distribution, because the points on the Normal Q-Q plot follow the diagonal reference line. In particular,

- there is no substantial curve (such as we'd see with data that were skewed)
- there is no particularly surprising behavior (curves away from the line) at either tail, so there's no obvious problem with outliers

10.8.2 Skew is indicated by monotonic curves in the Normal Q-Q plot

Data that come from a skewed distribution appear to curve away from a straight line in the Q-Q plot.

```

set.seed(123431) # so the results can be replicated

# simulate 200 observations from a beta(5, 2) distribution into the e1 variable
# simulate 200 observations from a beta(1, 5) distribution into the e2 variable
temp.2 <- data.frame(e1 = rbeta(200, 5, 2), e2 = rbeta(200, 1, 5))

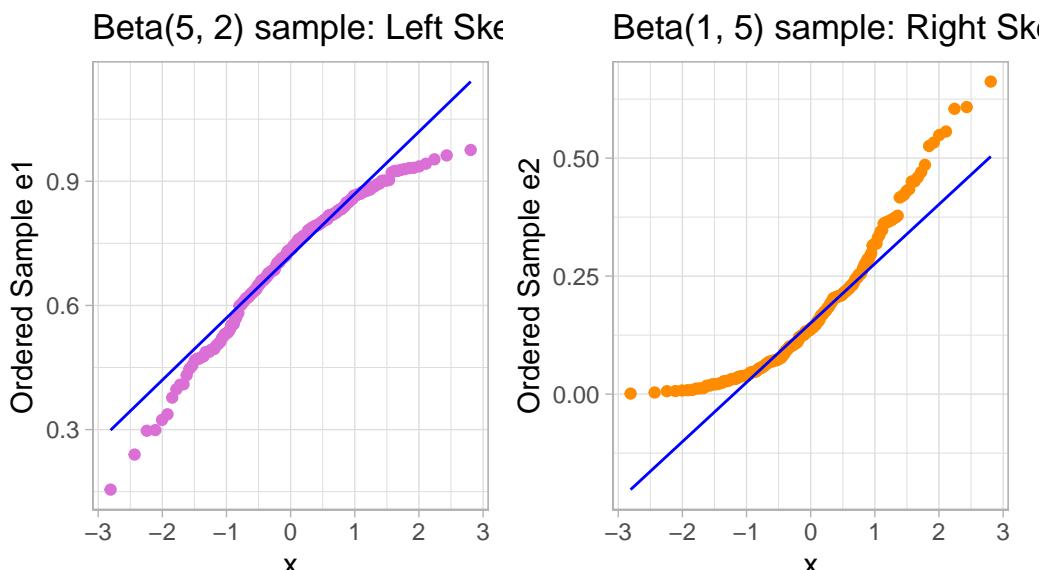
p1 <- ggplot(temp.2, aes(sample = e1)) +
  geom_qq(col = "orchid") + geom_qq_line(col = "blue") +
  theme_light() +
  labs(y = "Ordered Sample e1",
       title = "Beta(5, 2) sample: Left Skewed")

p2 <- ggplot(temp.2, aes(sample = e2)) +
  geom_qq(col = "darkorange") + geom_qq_line(col = "blue") +
  theme_light() +
  labs(y = "Ordered Sample e2",
       title = "Beta(1, 5) sample: Right Skewed")

p1 + p2 + plot_annotation(title = "200 observations from simulated Beta distributions")

```

200 observations from simulated Beta distributions



Note the bends away from a straight line in each sample. The non-Normality may be easier to see in a histogram.

```

res1 <- mosaic::favstats(~ e1, data = temp.2)
bin_w1 <- 0.025 # specify binwidth

p1 <- ggplot(temp.2, aes(x = e1)) +
  geom_histogram(binwidth = bin_w1,
                 fill = "orchid",
                 col = "black") +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res1$mean,
                            sd = res1$sd) *
      res1$n * bin_w1,
    col = "blue", size = 1.5) +
  labs(x = "Sample e1", y = "",
       title = "Beta(5,2) sample: Left Skew")

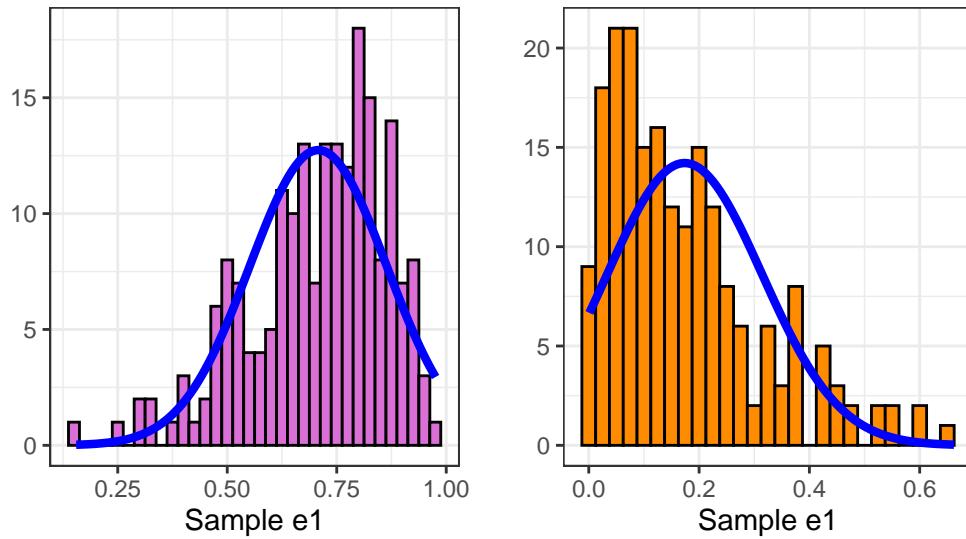
res2 <- mosaic::favstats(~ e2, data = temp.2)
bin_w2 <- 0.025 # specify binwidth

p2 <- ggplot(temp.2, aes(x = e2)) +
  geom_histogram(binwidth = bin_w2,
                 fill = "darkorange",
                 col = "black") +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res2$mean,
                            sd = res2$sd) *
      res2$n * bin_w2,
    col = "blue", size = 1.5) +
  labs(x = "Sample e1", y = "",
       title = "Beta(1,5) sample: Right Skew")

p1 + p2 + plot_annotation(caption = "Histograms with Normal curve superimposed")

```

Beta(5,2) sample: Left Skew Beta(1,5) sample: Right Skew



Histograms with Normal curve superimposed

10.8.3 Direction of Skew

In each of these pairs of plots, we see the same basic result.

- The left plot (for data e1) shows left skew, with a longer tail on the left hand side and more clustered data at the right end of the distribution.
- The right plot (for data e2) shows right skew, with a longer tail on the right hand side, the mean larger than the median, and more clustered data at the left end of the distribution.

10.8.4 Outlier-proneness is indicated by “s-shaped” curves in a Normal Q-Q plot

- Heavy-tailed but symmetric distributions are indicated by reverse “S”-shapes, as shown on the left below.
- Light-tailed but symmetric distributions are indicated by “S” shapes in the plot, as shown on the right below.

```
set.seed(4311) # so the results can be replicated

# sample 200 observations from each of two probability distributions
temp.3 <- data.frame(s1 = rcauchy(200, location=10, scale = 1),
                      s2 = runif(200, -30, 30))
```

```

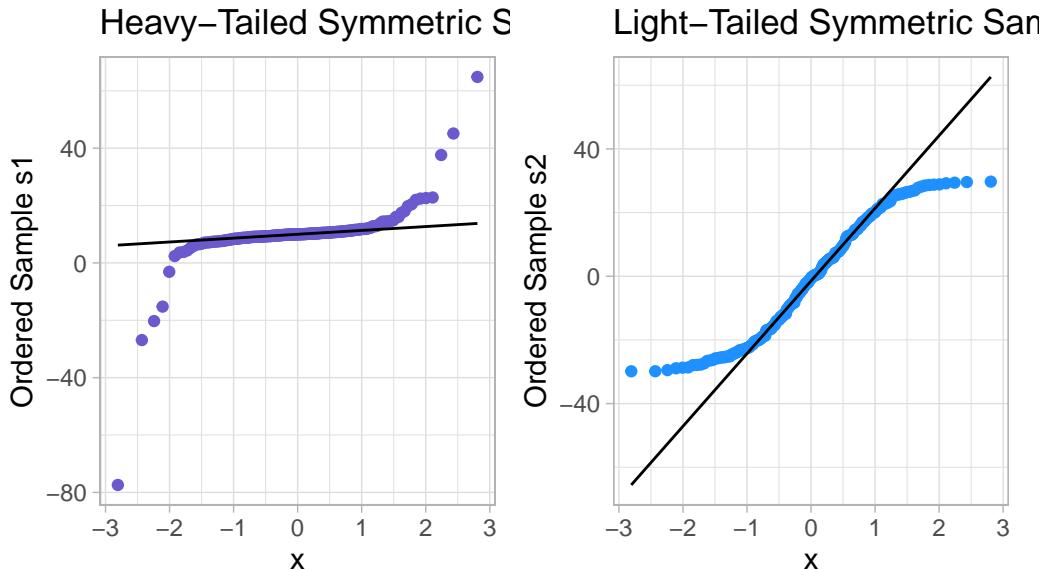
p1 <- ggplot(temp.3, aes(sample = s1)) +
  geom_qq(col = "slateblue") + geom_qq_line(col = "black") +
  theme_light() +
  labs(y = "Ordered Sample s1",
       title = "Heavy-Tailed Symmetric Sample s1")

p2 <- ggplot(temp.3, aes(sample = s2)) +
  geom_qq(col = "dodgerblue") + geom_qq_line(col = "black") +
  theme_light() +
  labs(y = "Ordered Sample s2",
       title = "Light-Tailed Symmetric Sample s2")

p1 + p2 + plot_annotation(title = "200 observations from simulated distributions")

```

200 observations from simulated distributions



And, we can also visualize these simulations with histograms, although they're less helpful for understanding tail behavior than they are for skew.

```

res1 <- mosaic::favstats(~ s1, data = temp.3)
bin_w1 <- 20 # specify binwidth

```

```

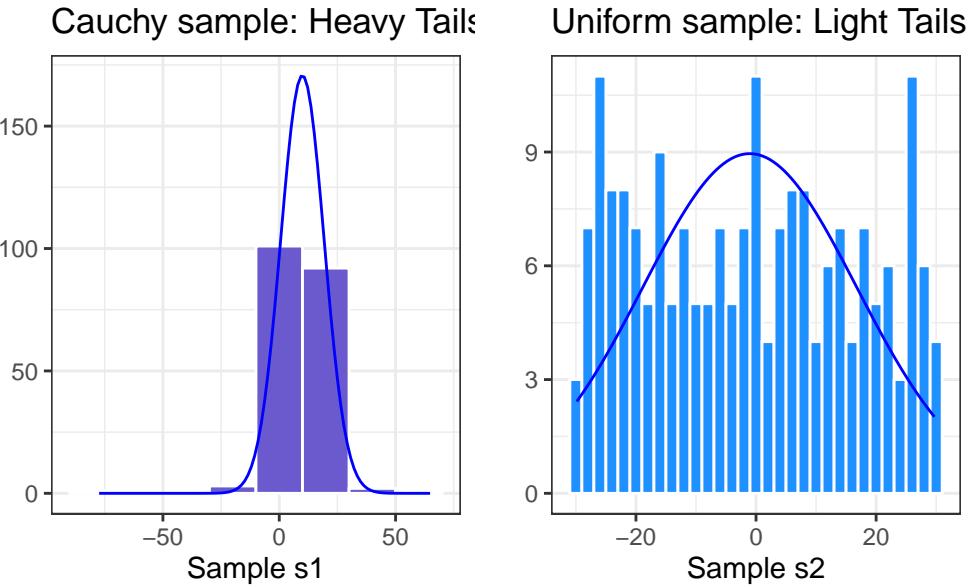
p1 <- ggplot(temp.3, aes(x = s1)) +
  geom_histogram(binwidth = bin_w1,
                 fill = "slateblue",
                 col = "white") +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res1$mean,
                            sd = res1$sd) *
    res1$n * bin_w1,
    col = "blue") +
  labs(x = "Sample s1", y = "",
       title = "Cauchy sample: Heavy Tails")

res2 <- mosaic::favstats(~ s2, data = temp.3)
bin_w2 <- 2 # specify binwidth

p2 <- ggplot(temp.3, aes(x = s2)) +
  geom_histogram(binwidth = bin_w2,
                 fill = "dodgerblue",
                 col = "white") +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res2$mean,
                            sd = res2$sd) *
    res2$n * bin_w2,
    col = "blue") +
  labs(x = "Sample s2", y = "",
       title = "Uniform sample: Light Tails")

p1 + p2 + plot_annotation(caption = "Histograms with Normal curve superimposed")

```



Histograms with Normal curve superimposed

Instead, boxplots (here augmented with violin plots) can be more helpful when thinking about light-tailed vs. heavy-tailed distributions.

```

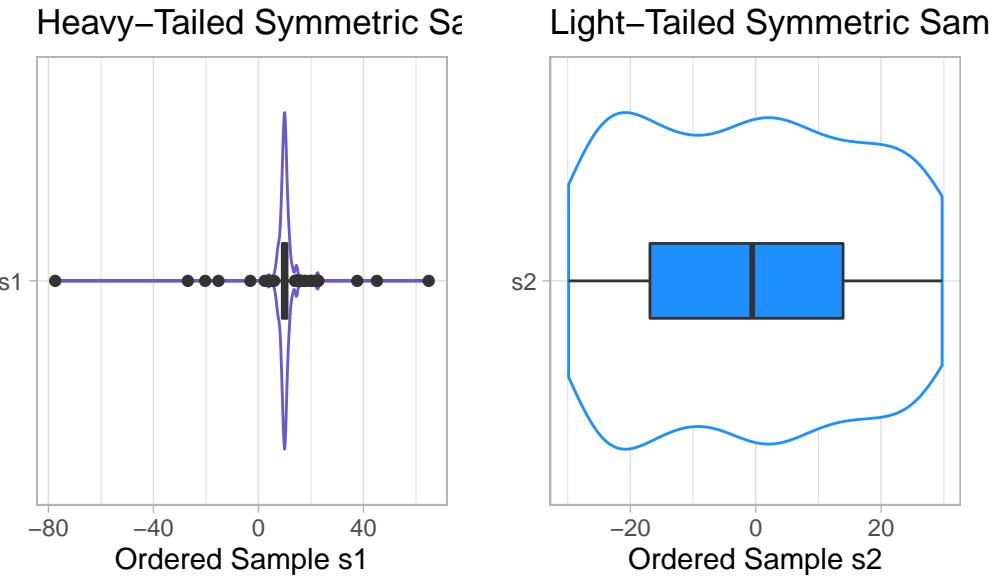
p1 <- ggplot(temp.3, aes(x = "s1", y = s1)) +
  geom_violin(col = "slateblue") +
  geom_boxplot(fill = "slateblue", width = 0.2) +
  theme_light() +
  coord_flip() +
  labs(y = "Ordered Sample s1", x = "",
       title = "Heavy-Tailed Symmetric Sample s1")

p2 <- ggplot(temp.3, aes(x = "s2", y = s2)) +
  geom_violin(col = "dodgerblue") +
  geom_boxplot(fill = "dodgerblue", width = 0.2) +
  theme_light() +
  coord_flip() +
  labs(y = "Ordered Sample s2", x = "",
       title = "Light-Tailed Symmetric Sample s2")

p1 + p2 + plot_annotation(title = "200 observations from simulated distributions")

```

200 observations from simulated distributions



```
rm(temp.1, temp.2, temp.3, p1, p2, res, res1, res2, bin_w, bin_w1, bin_w2) # cleaning up
```

10.9 Can a Normal Distribution Fit the nnyfs energy data Well?

The `energy` data we've been studying shows meaningful signs of right skew.

```
p1 <- ggplot(nnyfs, aes(sample = energy)) +
  geom_qq(col = "coral", size = 2) +
  geom_qq_line(col = "blue") +
  theme_light() +
  labs(title = "Energy Consumed",
       y = "Sorted Energy data")

res <- mosaic::favstats(~ energy, data = nnyfs)
bin_w <- 250 # specify binwidth

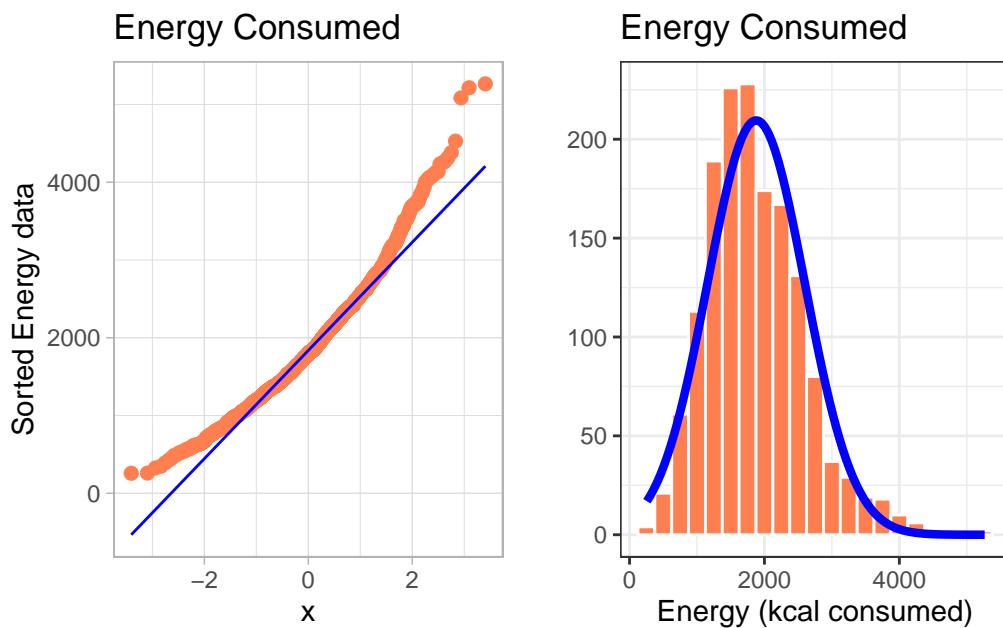
p2 <- ggplot(nnyfs, aes(x = energy)) +
  geom_histogram(binwidth = bin_w,
                 fill = "coral",
                 col = "white") +
```

```

theme_bw() +
stat_function(
  fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) *
  res$n * bin_w,
  col = "blue", size = 1.5) +
labs(x = "Energy (kcal consumed)", y = "",
     title = "Energy Consumed")

```

p1 + p2



- Skewness is indicated by the curve in the Normal Q-Q plot. Curving up and away from the line in both tails suggests right skew, as does the histogram.

What if we plotted not the original `energy` values (all of which are positive) but instead plotted the square roots of the `energy` values?

- Compare these two plots - the left describes the distribution of the original energy data from the NNYFS data frame, and the right plot shows the distribution of the square root of those values.

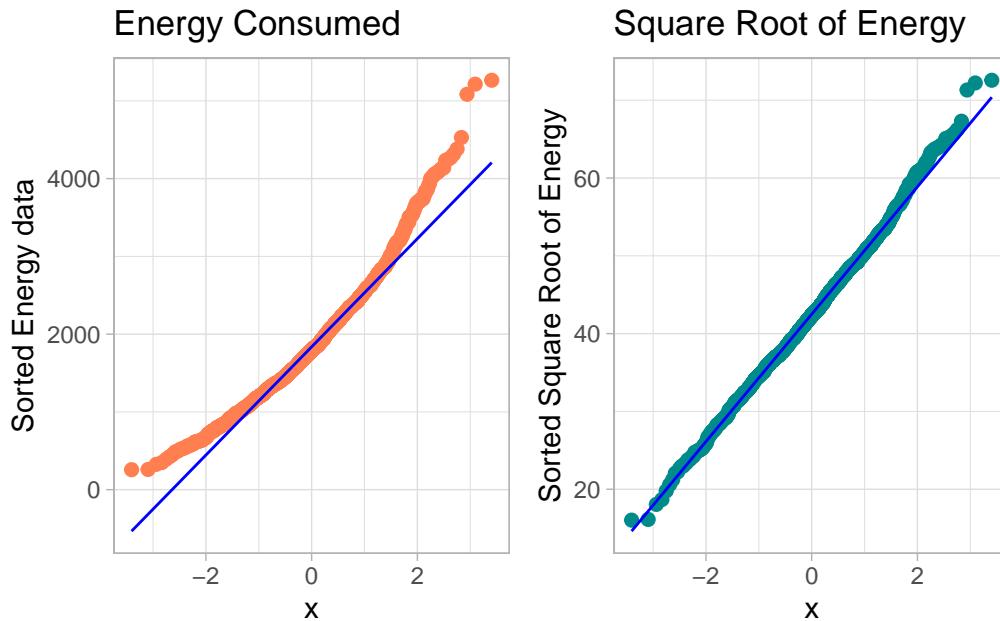
```

p1 <- ggplot(nnyfs, aes(sample = energy)) +
  geom_qq(col = "coral", size = 2) +
  geom_qq_line(col = "blue") +
  theme_light() +
  labs(title = "Energy Consumed",
       y = "Sorted Energy data")

p2 <- ggplot(nnyfs, aes(sample = sqrt(energy))) +
  geom_qq(col = "darkcyan", size = 2) +
  geom_qq_line(col = "blue") +
  theme_light() +
  labs(title = "Square Root of Energy",
       y = "Sorted Square Root of Energy")

p1 + p2

```



- The left plot shows substantial **right** or *positive* skew
- The right plot shows there's much less skew after the square root has been taken.

Our conclusion is that a Normal model is a far better fit to the square root of the energy values than it is to the raw energy values.

The effect of taking the square root may be clearer from the histograms below, with Normal

models superimposed.

```
res <- mosaic::favstats(~ energy, data = nnyfs)
bin_w <- 250 # specify binwidth

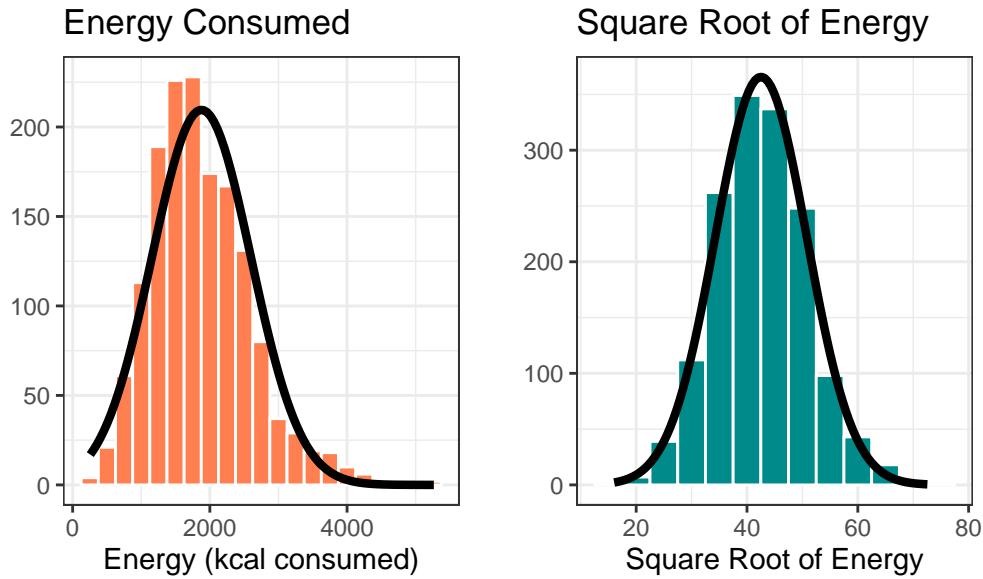
p1 <- ggplot(nnyfs, aes(x = energy)) +
  geom_histogram(binwidth = bin_w,
                 fill = "coral",
                 col = "white") +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) *
      res$n * bin_w,
    col = "black", size = 1.5) +
  labs(x = "Energy (kcal consumed)", y = "",
       title = "Energy Consumed")

res2 <- mosaic::favstats(~ sqrt(energy), data = nnyfs)
bin_w2 <- 5 # specify binwidth

p2 <- ggplot(nnyfs, aes(x = sqrt(energy))) +
  geom_histogram(binwidth = bin_w2,
                 fill = "darkcyan",
                 col = "white") +
  theme_bw() +
  stat_function(
    fun = function(x) dnorm(x, mean = res2$mean,
                           sd = res2$sd) *
      res2$n * bin_w2,
    col = "black", size = 1.5) +
  labs(x = "Square Root of Energy", y = "",
       title = "Square Root of Energy")

p1 + p2 + plot_annotation(title = "Comparing energy to sqrt(energy)")
```

Comparing energy to $\text{sqrt}(\text{energy})$



```
rm(p1, p2, bin_w, bin_w2, res, res2) # cleanup
```

When we are confronted with a variable that is not Normally distributed but that we wish was Normally distributed, it is sometimes useful to consider whether working with a **transformation** of the data will yield a more helpful result, as the square root does in this instance.

The rest of this Chapter provides some guidance about choosing from a class of power transformations that can reduce the impact of non-Normality in unimodal data.

- When we are confronted with a variable that is not Normally distributed but that we wish was Normally distributed, it is sometimes useful to consider whether working with a transformation of the data will yield a more helpful result.
- Many statistical methods, including t tests and analyses of variance, assume Normal distributions.
- We'll discuss using R to assess a range of what are called Box-Cox power transformations, via plots, mainly.

10.10 The Ladder of Power Transformations

The key notion in re-expression of a single variable to obtain a distribution better approximated by the Normal or re-expression of an outcome in a simple regression model is that of a **ladder of power transformations**, which applies to any unimodal data.

Power	Transformation
3	x^3
2	x^2
1	x (unchanged)
0.5	$x^{0.5} = \sqrt{x}$
0	$\ln x$
-0.5	$x^{-0.5} = 1/\sqrt{x}$
-1	$x^{-1} = 1/x$
-2	$x^{-2} = 1/x^2$

10.11 Using the Ladder

As we move further away from the *identity* function (power = 1) we change the shape more and more in the same general direction.

- For instance, if we try a logarithm, and this seems like too much of a change, we might try a square root instead.
- Note that this ladder (which like many other things is due to John Tukey) uses the logarithm for the “power zero” transformation rather than the constant, which is what x^0 actually is.
- If the variable x can take on negative values, we might take a different approach. If x is a count of something that could be zero, we often simply add 1 to x before transformation.

The ladder of power transformations is particularly helpful when we are confronted with data that shows skew.

- To handle right skew (where the mean exceeds the median) we usually apply powers below 1.
- To handle left skew (where the median exceeds the mean) we usually apply powers greater than 1.

The most common transformations are the square (power 2), the square root (power 1/2), the logarithm (power 0) and the inverse (power -1), and I usually restrict myself to those options in practical work.

10.12 Protein Consumption in the NNYFS data

Here are the protein consumption (in grams) results from the NNYFS data.

```

mosaic::favstats(~ protein, data = nnyfs)

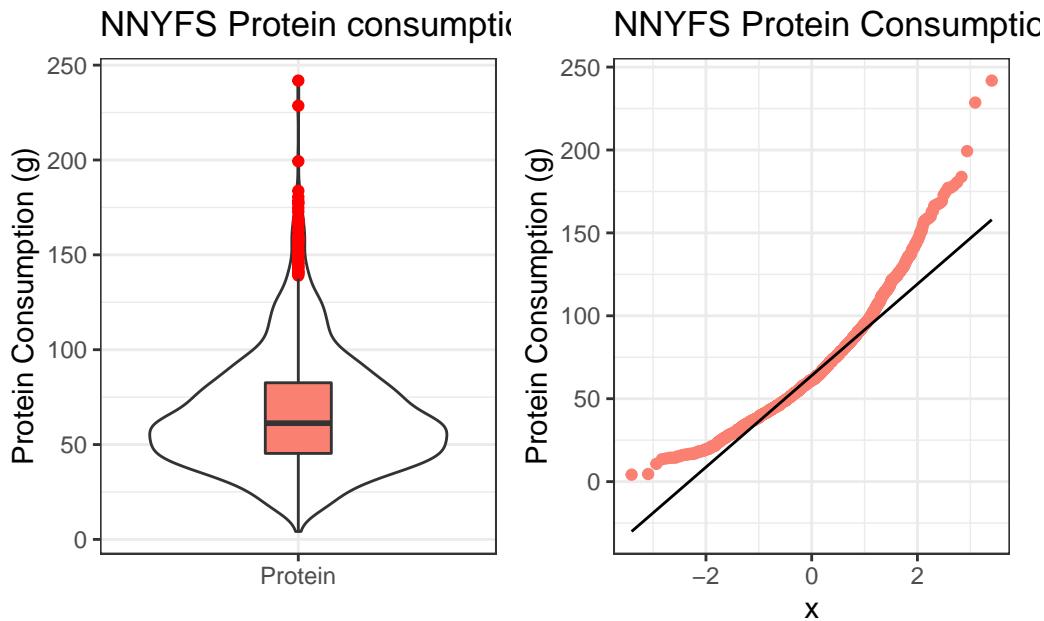
min      Q1 median      Q3    max      mean       sd     n missing
4.18 45.33 61.255 82.565 241.84 66.90148 30.96319 1518          0

p1 <- ggplot(nnyfs, aes(x = "Protein", y = protein)) +
  geom_violin() +
  geom_boxplot(width = 0.2, fill = "salmon",
                outlier.color = "red") +
  labs(title = "NNYFS Protein consumption",
       x = "", y = "Protein Consumption (g)")

p2 <- ggplot(nnyfs, aes(sample = protein)) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "NNYFS Protein Consumption",
       y = "Protein Consumption (g)")

p1 + p2

```



The key point here is that we see several signs of meaningful right skew, and we'll want to consider a transformation that might make a Normal model more plausible.

10.12.1 Using patchwork to compose plots

As we mentioned previously, I feel that the slickest approach to composing how a series of plots are placed together is available in the `patchwork` package. Here's another example.

```
res <- mosaic::favstats(~ protein, data = nnyfs)
bin_w <- 5 # specify binwidth

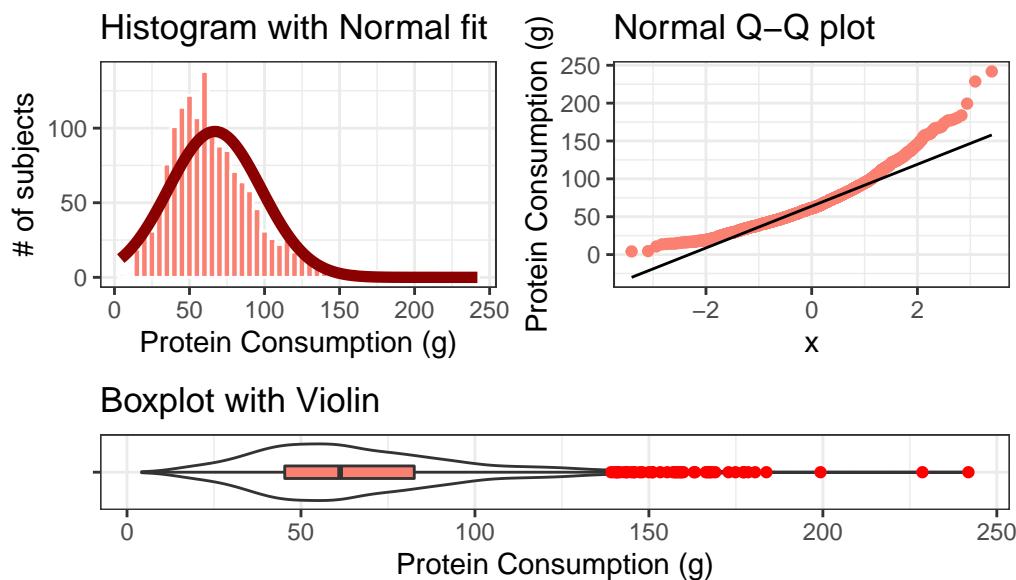
p1 <- ggplot(nnyfs, aes(x = protein)) +
  geom_histogram(binwidth = bin_w,
                 fill = "salmon",
                 col = "white") +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) *
      res$n * bin_w,
    col = "darkred", size = 2) +
  labs(title = "Histogram with Normal fit",
       x = "Protein Consumption (g)", y = "# of subjects")

p2 <- ggplot(nnyfs, aes(sample = protein)) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "Normal Q-Q plot",
       y = "Protein Consumption (g)")

p3 <- ggplot(nnyfs, aes(x = "", y = protein)) +
  geom_violin() +
  geom_boxplot(width = 0.2, fill = "salmon",
               outlier.color = "red") +
  coord_flip() +
  labs(title = "Boxplot with Violin",
       x = "", y = "Protein Consumption (g)")
```

```
p1 + p2 - p3 + plot_layout(ncol = 1, height = c(3, 1)) +
  plot_annotation(title = "NNYFS Protein Consumption")
```

NNYFS Protein Consumption



Again, the `patchwork` package repository at <https://patchwork.data-imaginist.com/index.html> has lots of nice examples to work from.

10.13 Can we transform the protein data?

As we've seen, the protein data are right skewed, and all of the values are strictly positive. If we want to use the tools of the Normal distribution to describe these data, we might try taking a step "down" our ladder from power 1 (raw data) to lower powers.

10.13.1 The Square Root

Would a square root applied to the protein data help alleviate that right skew?

```
res <- mosaic::favstats(~ sqrt(protein), data = nnyfs)
bin_w <- 1 # specify binwidth
```

```

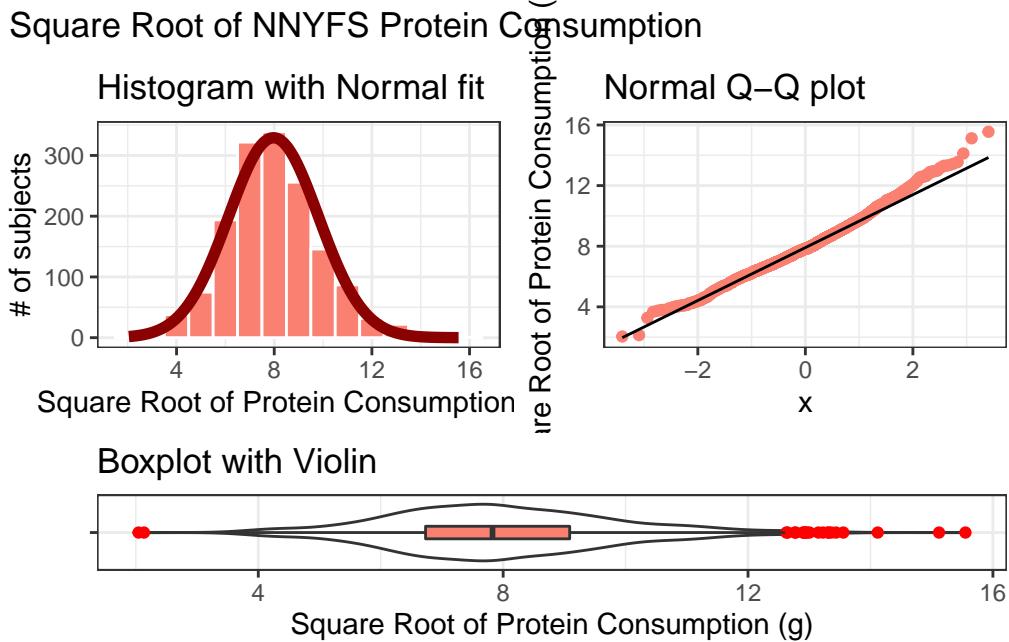
p1 <- ggplot(nnyfs, aes(x = sqrt(protein))) +
  geom_histogram(binwidth = bin_w,
                 fill = "salmon",
                 col = "white") +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                            sd = res$sd) *
      res$n * bin_w,
    col = "darkred", size = 2) +
  labs(title = "Histogram with Normal fit",
       x = "Square Root of Protein Consumption (g)", y = "# of subjects")

p2 <- ggplot(nnyfs, aes(sample = sqrt(protein))) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "Normal Q-Q plot",
       y = "Square Root of Protein Consumption (g)")

p3 <- ggplot(nnyfs, aes(x = "", y = sqrt(protein))) +
  geom_violin() +
  geom_boxplot(width = 0.2, fill = "salmon",
               outlier.color = "red") +
  coord_flip() +
  labs(title = "Boxplot with Violin",
       x = "", y = "Square Root of Protein Consumption (g)")

p1 + p2 - p3 + plot_layout(ncol = 1, height = c(3, 1)) +
  plot_annotation(title = "Square Root of NNYFS Protein Consumption")

```



That looks like a more symmetric distribution, certainly, although we still have some outliers on the right side of the distribution. Should we take another step down the ladder?

10.13.2 The Logarithm

We might also try a logarithm of the energy circumference data. We can use either the natural logarithm (`log`, in R) or the base-10 logarithm (`log10`, in R) - either will have the same impact on skew.

```
res <- mosaic::favstats(~ log(protein), data = nnyfs)
bin_w <- 0.5 # specify binwidth

p1 <- ggplot(nnyfs, aes(x = log(protein))) +
  geom_histogram(binwidth = bin_w,
                 fill = "salmon",
                 col = "white") +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) *
      res$n * bin_w,
    col = "darkred", size = 2) +
```

```

  labs(title = "Histogram with Normal fit",
       x = "Log of Protein Consumption (g)", y = "# of subjects")

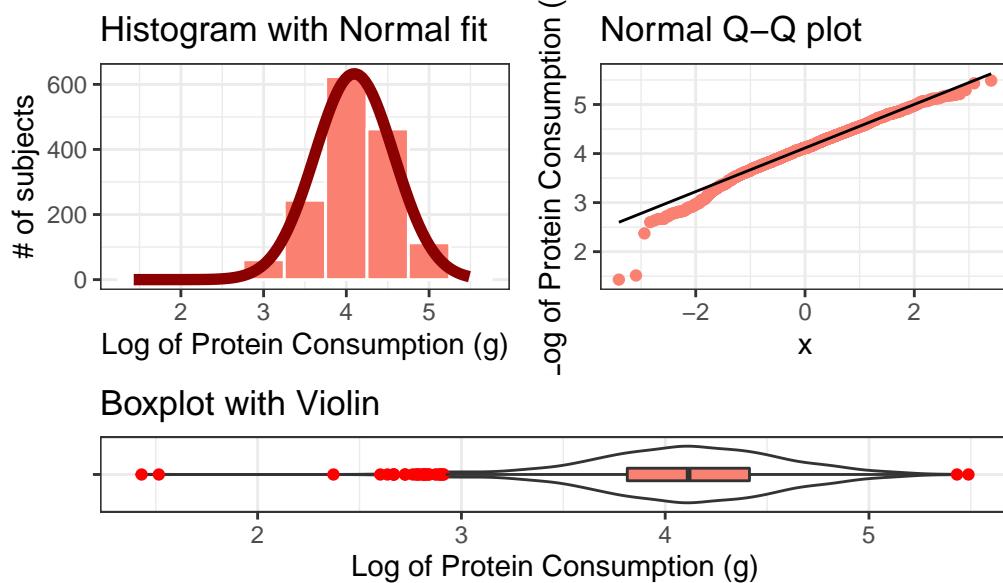
p2 <- ggplot(nnyfs, aes(sample = log(protein))) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "Normal Q-Q plot",
       y = "Log of Protein Consumption (g)")

p3 <- ggplot(nnyfs, aes(x = "", y = log(protein))) +
  geom_violin() +
  geom_boxplot(width = 0.2, fill = "salmon",
               outlier.color = "red") +
  coord_flip() +
  labs(title = "Boxplot with Violin",
       x = "", y = "Log of Protein Consumption (g)")

p1 + p2 - p3 + plot_layout(ncol = 1, height = c(3, 1)) +
  plot_annotation(title = "Logarithm of NNYFS Protein Consumption")

```

Logarithm of NNYFS Protein Consumption



Now, it looks like we may have gone too far in the other direction. It looks like the square root

is a sensible choice to try to improve the fit of a Normal model to the protein consumption data.

10.13.3 This course uses Natural Logarithms, unless otherwise specified

In this course, we will assume the use of natural logarithms unless we specify otherwise. Following R's convention, we will use `log` for natural logarithms.

10.14 What if we considered all 9 available transformations?

```
p1 <- ggplot(nnyfs, aes(sample = protein^3)) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "Cube (power 3)",
       y = "Protein, Cubed")

p2 <- ggplot(nnyfs, aes(sample = protein^2)) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "Square (power 2)",
       y = "Protein, Squared")

p3 <- ggplot(nnyfs, aes(sample = protein)) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "Original Data",
       y = "Protein (g)")

p4 <- ggplot(nnyfs, aes(sample = sqrt(protein))) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "sqrt (power 0.5)",
       y = "Square Root of Protein")

p5 <- ggplot(nnyfs, aes(sample = log(protein))) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "log (power 0)",
       y = "Natural Log of Protein")
```

```

p6 <- ggplot(nnyfs, aes(sample = protein^(-0.5))) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "1/sqrt (power -0.5)",
       y = "1/Square Root(Protein)")

p7 <- ggplot(nnyfs, aes(sample = 1/protein)) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "Inverse (power -1)",
       y = "1/Protein")

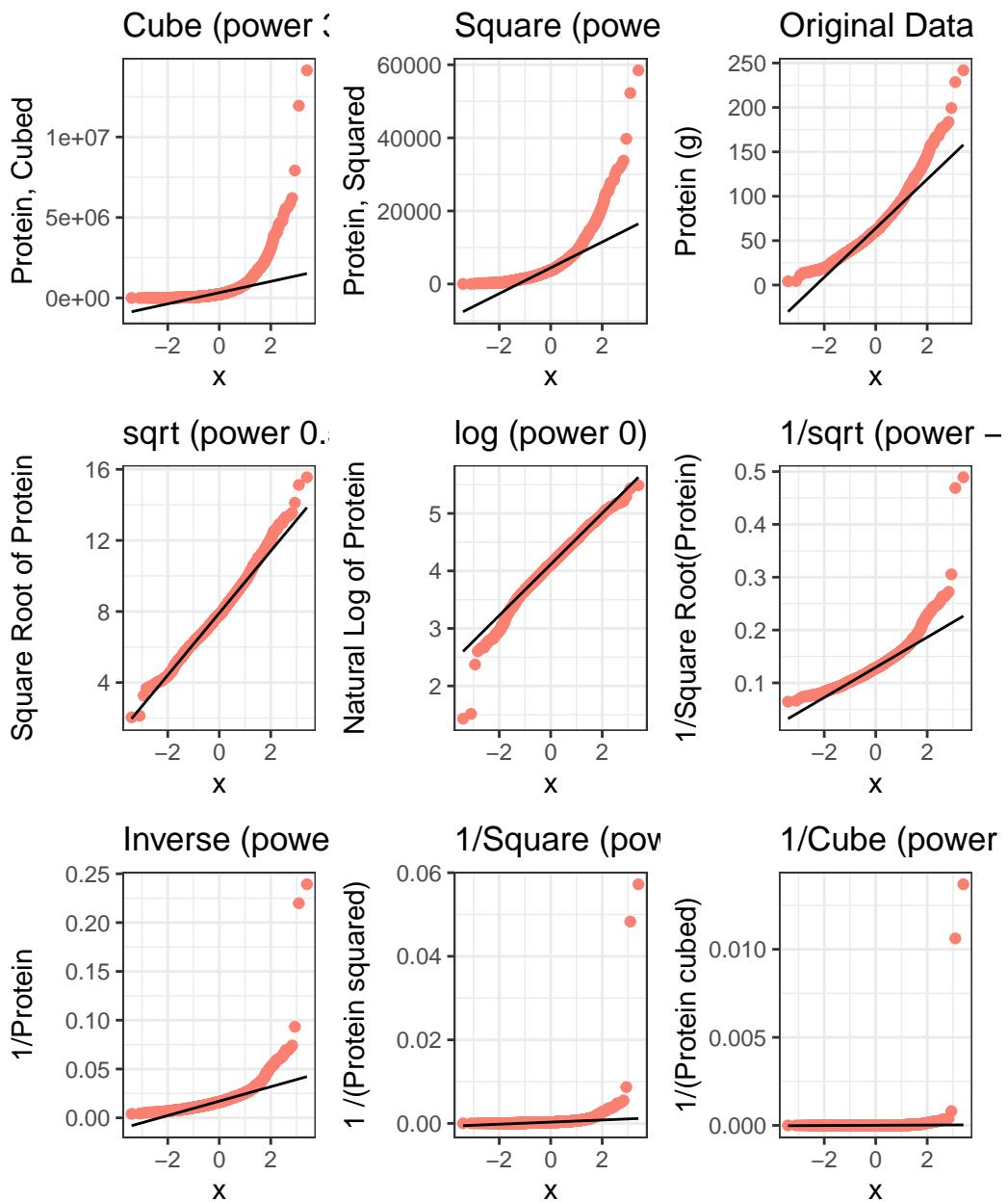
p8 <- ggplot(nnyfs, aes(sample = 1/(protein^2))) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "1/Square (power -2)",
       y = "1 /(Protein squared)")

p9 <- ggplot(nnyfs, aes(sample = 1/(protein^3))) +
  geom_qq(col = "salmon") +
  geom_qq_line(col = "black") +
  labs(title = "1/Cube (power -3)",
       y = "1/(Protein cubed)")

p1 + p2 + p3 + p4 + p5 + p6 + p7 + p8 + p9 +
  plot_layout(nrow = 3) +
  plot_annotation(title = "Transformations of NNYFS Protein Consumption")

```

Transformations of NNYFS Protein Consumption



The square root still appears to be the best choice of transformation here, even after we consider all 8 transformation of the raw data.

10.15 A Simulated Data Set

```
set.seed(431);
data2 <-
  data_frame(sample2 = 100*rbeta(n = 125, shape1 = 5, shape2 = 2))

Warning: `data_frame()` was deprecated in tibble 1.1.0.
Please use `tibble()` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

If we'd like to transform these data so as to better approximate a Normal distribution, where should we start? What transformation do you suggest?

```
res <- mosaic::favstats(~ sample2, data = data2)
bin_w <- 4 # specify binwidth

p1 <- ggplot(data2, aes(x = sample2)) +
  geom_histogram(binwidth = bin_w,
                 fill = "royalblue",
                 col = "white") +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) *
      res$n * bin_w,
    col = "darkred", size = 2) +
  labs(title = "Histogram with Normal fit",
       x = "Simulated Data", y = "# of subjects")

p2 <- ggplot(data2, aes(sample = sample2)) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "black") +
  labs(title = "Normal Q-Q plot",
       y = "Simulated Data")

p3 <- ggplot(data2, aes(x = "", y = sample2)) +
  geom_violin() +
  geom_boxplot(width = 0.3, fill = "royalblue",
               outlier.color = "royalblue") +
  coord_flip()
```

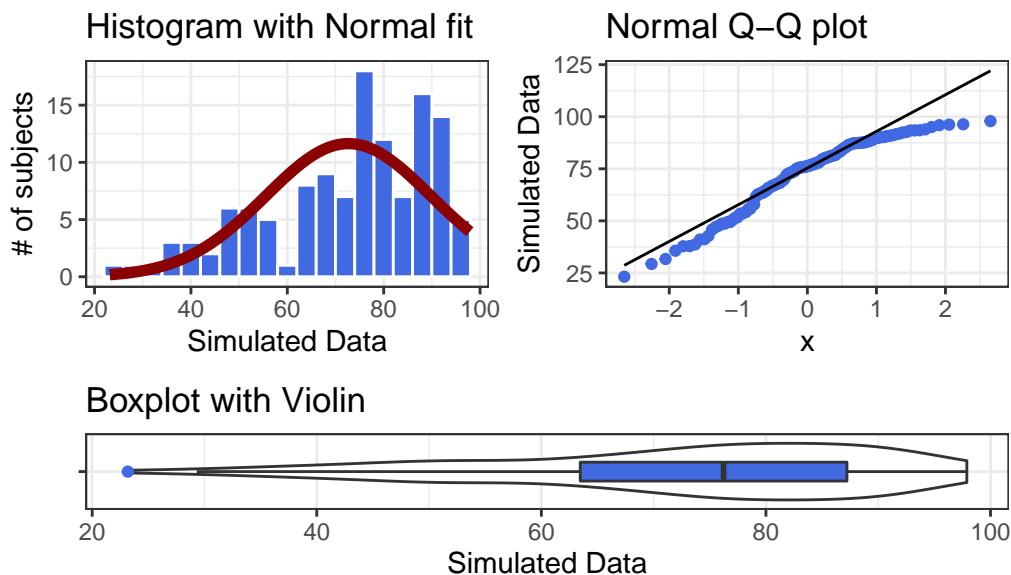
```

  labs(title = "Boxplot with Violin",
       x = "", y = "Simulated Data")

p1 + p2 - p3 + plot_layout(ncol = 1, height = c(3, 1)) +
  plot_annotation(title = "Simulated Data")

```

Simulated Data



Given the left skew in the data, it looks like a step up in the ladder is warranted, perhaps by looking at the square of the data?

```

res <- mosaic::favstats(~ sample2^2, data = data2)
bin_w <- 600 # specify binwidth

p1 <- ggplot(data2, aes(x = sample2^2)) +
  geom_histogram(binwidth = bin_w,
                 fill = "royalblue",
                 col = "white") +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) *
      res$n * bin_w,
    col = "darkred", size = 2) +

```

```

  labs(title = "Histogram with Normal fit",
       x = "Squared Simulated Data", y = "# of subjects")

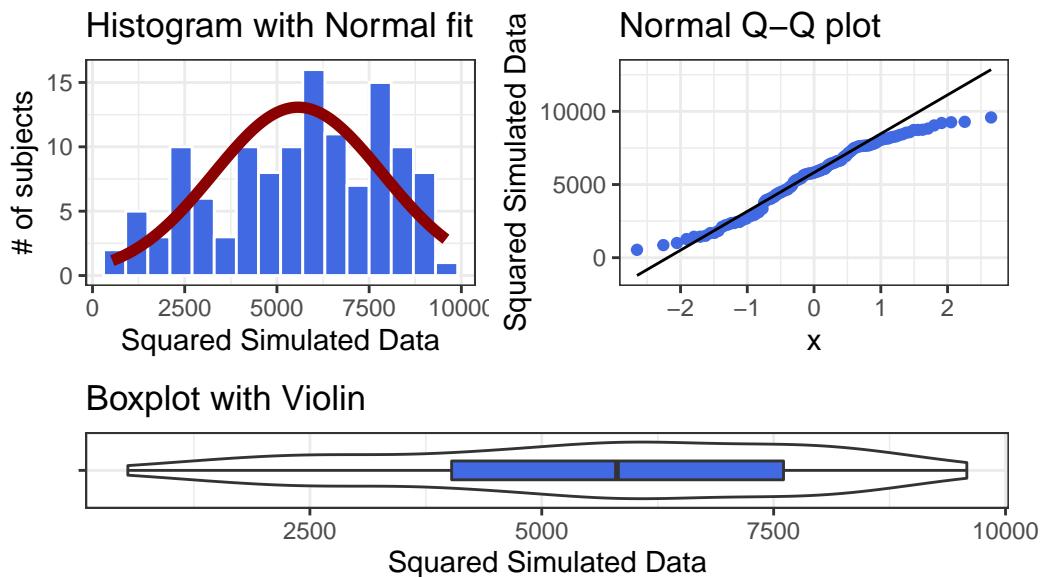
p2 <- ggplot(data2, aes(sample = sample2^2)) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "black") +
  labs(title = "Normal Q-Q plot",
       y = "Squared Simulated Data")

p3 <- ggplot(data2, aes(x = "", y = sample2^2)) +
  geom_violin() +
  geom_boxplot(width = 0.3, fill = "royalblue",
               outlier.color = "royalblue") +
  coord_flip() +
  labs(title = "Boxplot with Violin",
       x = "", y = "Squared Simulated Data")

p1 + p2 - p3 + plot_layout(ncol = 1, height = c(3, 1)) +
  plot_annotation(title = "Squared Simulated Data")

```

Squared Simulated Data



Looks like at best a modest improvement. How about cubing the data, instead?

```

res <- mosaic::favstats(~ sample2^3, data = data2)
bin_w <- 100000 # specify binwidth

p1 <- ggplot(data2, aes(x = sample2^3)) +
  geom_histogram(binwidth = bin_w,
                 fill = "royalblue",
                 col = "white") +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,
                            sd = res$sd) *
      res$n * bin_w,
    col = "darkred", size = 2) +
  labs(title = "Histogram with Normal fit",
       x = "Cubed Simulated Data", y = "# of subjects")

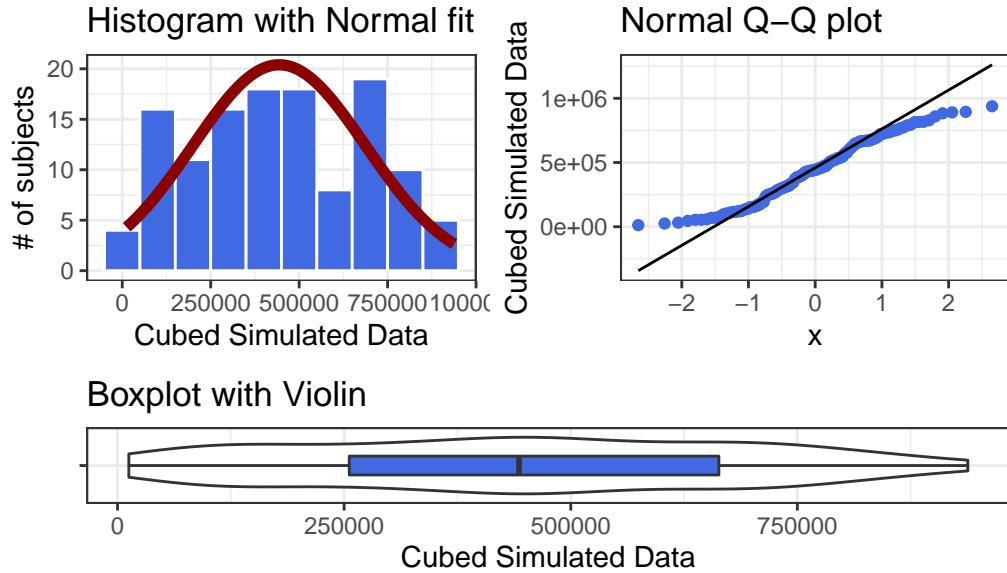
p2 <- ggplot(data2, aes(sample = sample2^3)) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "black") +
  labs(title = "Normal Q-Q plot",
       y = "Cubed Simulated Data")

p3 <- ggplot(data2, aes(x = "", y = sample2^3)) +
  geom_violin() +
  geom_boxplot(width = 0.3, fill = "royalblue",
               outlier.color = "royalblue") +
  coord_flip() +
  labs(title = "Boxplot with Violin",
       x = "", y = "Cubed Simulated Data")

p1 + p2 - p3 + plot_layout(ncol = 1, height = c(3, 1)) +
  plot_annotation(title = "Cubed Simulated Data")

```

Cubed Simulated Data



The newly transformed (cube of the) data appears more symmetric, although somewhat light-tailed. Perhaps a Normal model would be more appropriate now, although the standard deviation is likely to overstate the variation we see in the data due to the light tails. Again, I wouldn't be thrilled using a cube in practical work, as it is so hard to interpret, but it does look like a reasonable choice here.

10.16 What if we considered all 9 available transformations?

```
p1 <- ggplot(data2, aes(sample = sample2^3)) +  
  geom_qq(col = "royalblue") +  
  geom_qq_line(col = "black") +  
  labs(title = "Cube (power 3)")  
  
p2 <- ggplot(data2, aes(sample = sample2^2)) +  
  geom_qq(col = "royalblue") +  
  geom_qq_line(col = "black") +  
  labs(title = "Square (power 2)")  
  
p3 <- ggplot(data2, aes(sample = sample2)) +  
  geom_qq(col = "royalblue") +
```

```

geom_qq_line(col = "black") +
  labs(title = "Original Data")

p4 <- ggplot(data2, aes(sample = sqrt(sample2))) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "black") +
  labs(title = "sqrt (power 0.5)")

p5 <- ggplot(data2, aes(sample = log(sample2))) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "black") +
  labs(title = "log (power 0)")

p6 <- ggplot(data2, aes(sample = sample2^(0.5))) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "black") +
  labs(title = "1/sqrt (power -0.5)")

p7 <- ggplot(data2, aes(sample = 1/sample2)) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "black") +
  labs(title = "Inverse (power -1)")

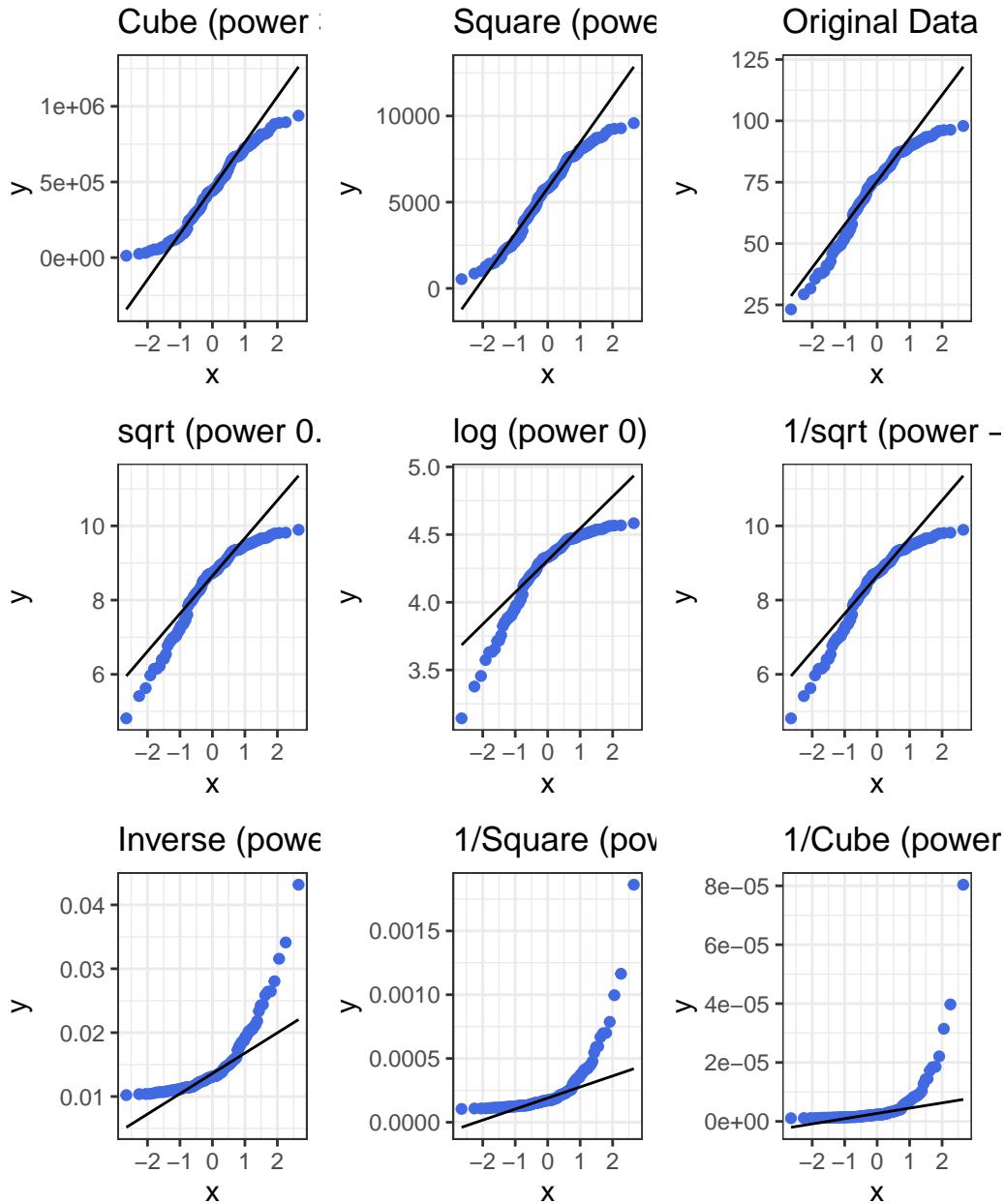
p8 <- ggplot(data2, aes(sample = 1/(sample2^2))) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "black") +
  labs(title = "1/Square (power -2)")

p9 <- ggplot(data2, aes(sample = 1/(sample2^3))) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "black") +
  labs(title = "1/Cube (power -3)")

p1 + p2 + p3 + p4 + p5 + p6 + p7 + p8 + p9 +
  plot_layout(nrow = 3) +
  plot_annotation(title = "Transformations of Simulated Sample")

```

Transformations of Simulated Sample



Again, either the cube or the square looks like best choice here, in terms of creating a more symmetric (albeit light-tailed) distribution.

11 Straight Line Models

11.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(broom)
library(knitr)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

11.2 Assessing A Scatterplot

Let's consider the relationship of `protein` and `fat` consumption for children in the `nnyfs` data.

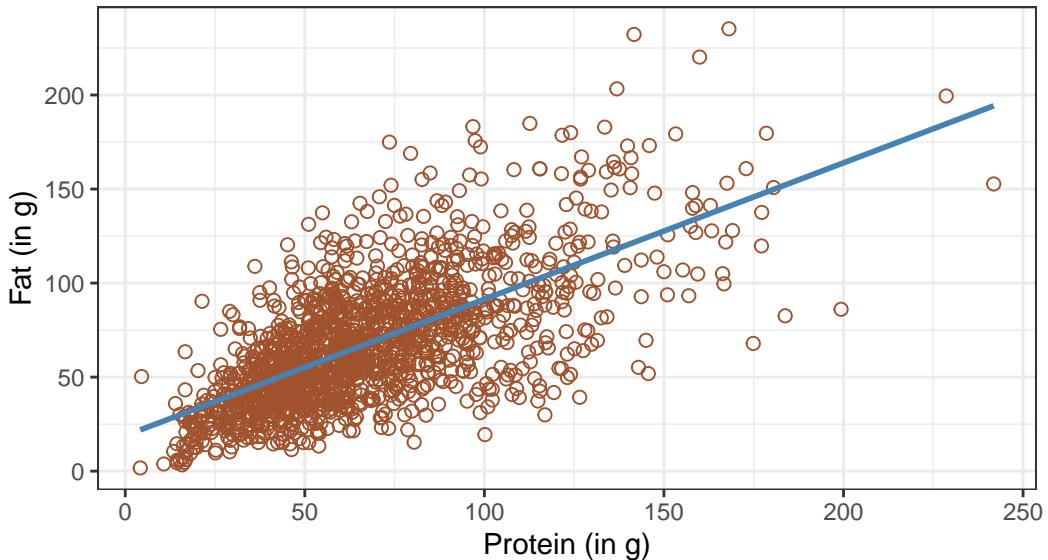
```
nnyfs <- read_rds("data/nnyfs.Rds")
```

We'll begin our investigation, as we always should, by drawing a relevant picture. For the association of two quantitative variables, a **scatterplot** is usually the right start. Each subject in the `nnyfs` data is represented by one of the points below. To the plot, I've also used `geom_smooth` to add a straight line regression model, which we'll discuss later.

```
ggplot(data = nnyfs, aes(x = protein, y = fat)) +
  geom_point(shape = 1, size = 2, col = "sienna") +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, col = "steelblue") +
  labs(title = "Protein vs. Fat consumption",
       subtitle = "Children in the NNYFS data, with fitted straight line regression model",
       x = "Protein (in g)", y = "Fat (in g)")
```

Protein vs. Fat consumption

Children in the NNYFS data, with fitted straight line regression model



Here, I've arbitrarily decided to place `fat` on the vertical axis, and `protein` on the horizontal. Fitting a prediction model to this scatterplot will then require that we predict `fat` on the basis of `protein`.

In this case, the pattern appears to be:

1. **direct**, or positive, in that the values of the x variable (`protein`) increase, so do the values of the y variable (`fat`). Essentially, it appears that subjects who consumed more protein also consumed more fat, but we don't know cause and effect here.
2. fairly **linear** in that most of the points cluster around what appears to be a pattern which is well-fitted by a straight line.
3. moderately **strong** in that the range of values for `fat` associated with any particular value of `protein` is fairly tight. If we know someone's protein consumption, that should meaningfully improve our ability to predict their fat consumption, among the subjects in these data.
4. that we see some unusual or **outlier** values, further away from the general pattern of most subjects shown in the data.

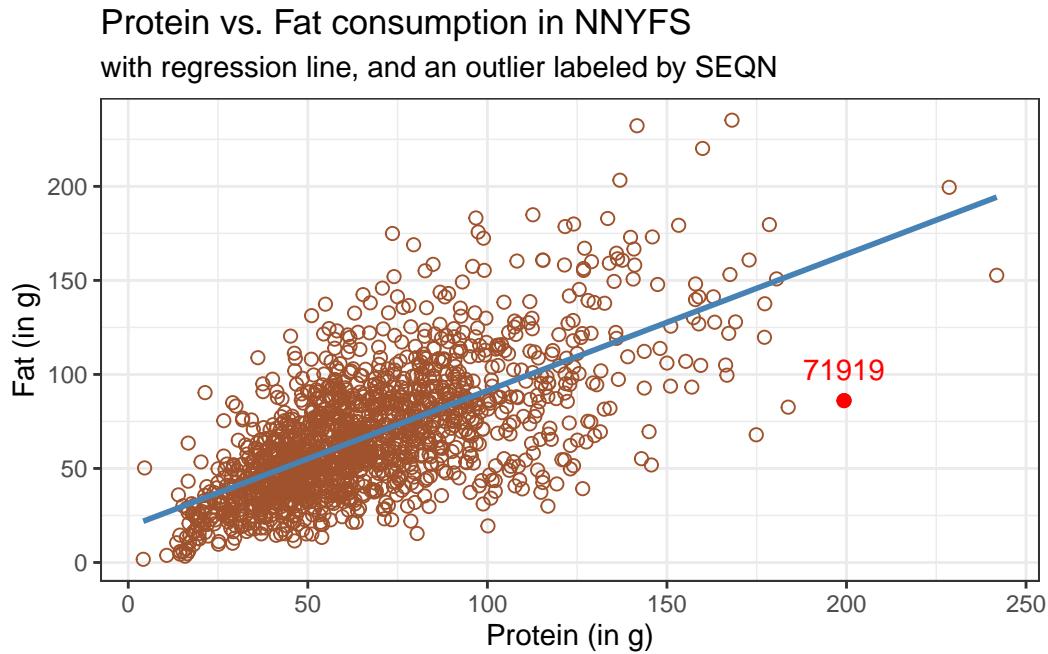
11.2.1 Highlighting an unusual point

Consider the subject with protein consumption close to 200 g, whose fat consumption is below 100 g. That's well below the prediction of the linear model for example. We can identify the subject because it is the only person with `protein` > 190 and `fat` < 100 with BMI > 35 and

`waist.circ < 70`. So I'll create a subset of the `nnyfs` data containing the point that meets that standard, and then add a red point and a label to the plot.

```
# identify outlier and place it in data frame s1
s1 <- filter(nnyfs, protein > 190 & fat < 100)

ggplot(data = nnyfs, aes(x = protein, y = fat)) +
  geom_point(shape = 1, size = 2, col = "sienna") +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x, col = "steelblue") +
  geom_point(data = s1, size = 2, col = "red") +
  geom_text(data = s1, label = s1$SEQN,
            vjust = -1, col = "red") +
  labs(title = "Protein vs. Fat consumption in NNYFS",
       subtitle = "with regression line, and an outlier labeled by SEQN",
       x = "Protein (in g)", y = "Fat (in g)")
```



While this subject is hardly the only unusual point in the data set, it is one of the more unusual ones, in terms of its vertical distance from the regression line. We can identify the subject by printing (part of) the tibble we created.

```
s1 |> select(SEQN, sex, race_eth, age_child, protein, fat) |> kable()
```

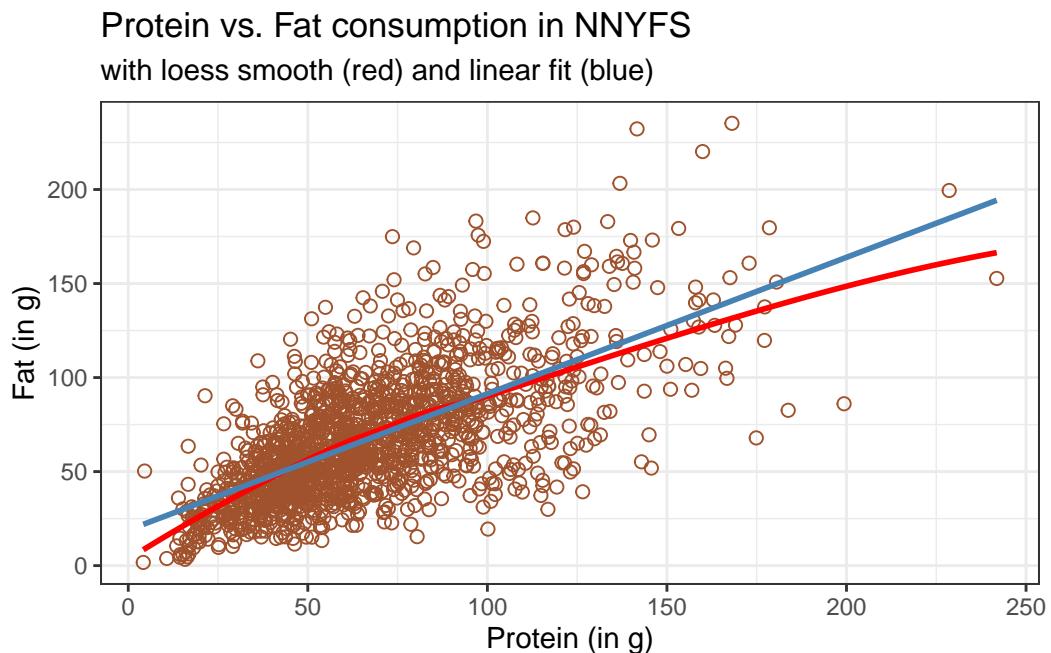
SEQN	sex	race_eth	age_child	protein	fat
71919	Female	2_White Non-Hispanic	14	199.33	86.08

Now, does it seem to you like a straight line model will describe this **protein-fat** relationship well?

11.2.2 Adding a Scatterplot Smooth using loess

Next, we'll use the **loess** procedure to fit a smooth curve to the data, which attempts to capture the general pattern.

```
ggplot(data = nnyfs, aes(x = protein, y = fat)) +
  geom_point(shape = 1, size = 2, col = "sienna") +
  geom_smooth(method = "loess", se = FALSE, formula = y ~ x, col = "red") +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x, col = "steelblue") +
  labs(title = "Protein vs. Fat consumption in NNYFS",
       subtitle = "with loess smooth (red) and linear fit (blue)",
       x = "Protein (in g)", y = "Fat (in g)")
```



This “loess” smooth curve is fairly close to the straight line fit, indicating that perhaps a linear regression model might fit the data well.

A **loess smooth** is a method of fitting a local polynomial regression model that R uses as its default smooth for scatterplots with fewer than 1000 observations. Think of the loess as a way of fitting a curve to data by tracking (at point x) the points within a neighborhood of point x , with more emphasis given to points near x . It can be adjusted by tweaking two specific parameters, in particular:

- a **span** parameter (defaults to 0.75) which is also called α in the literature, that controls the degree of smoothing (essentially, how large the neighborhood should be), and
- a **degree** parameter (defaults to 2) which specifies the degree of polynomial to be used. Normally, this is either 1 or 2 - more complex functions are rarely needed for simple scatterplot smoothing.

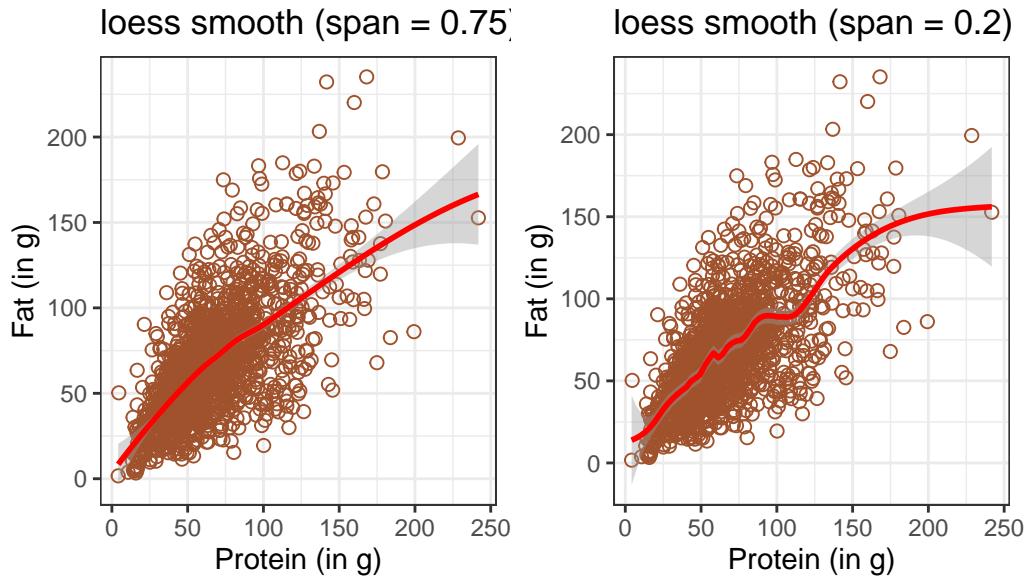
In addition to the curve, smoothing procedures can also provide confidence intervals around their main fitted line. Consider the following plot, which adjusts the span and also adds in the confidence intervals.

```
p1 <- ggplot(data = nnyfs, aes(x = protein, y = fat)) +
  geom_point(shape = 1, size = 2, col = "sienna") +
  geom_smooth(method = "loess", span = 0.75, se = TRUE,
              col = "red", formula = y ~ x) +
  labs(title = "loess smooth (span = 0.75)",
       x = "Protein (in g)", y = "Fat (in g)")

p2 <- ggplot(data = nnyfs, aes(x = protein, y = fat)) +
  geom_point(shape = 1, size = 2, col = "sienna") +
  geom_smooth(method = "loess", span = 0.2, se = TRUE,
              col = "red", formula = y ~ x) +
  labs(title = "loess smooth (span = 0.2)",
       x = "Protein (in g)", y = "Fat (in g)")

p1 + p2 +
  plot_annotation(title = "Impact of adjusting loess span: NNYFS")
```

Impact of adjusting loess smooth span: NNYFS



By reducing the size of the span, the plot on the right shows a somewhat less “smooth” function than the plot on the left.

11.2.3 What Line Does R Fit?

Returning to the linear regression model, how can we, mathematically, characterize that line? As with any straight line, our model equation requires us to specify two parameters: a slope and an intercept (sometimes called the y-intercept.)

To identify the equation R used to fit this line (using the method of least squares), we use the `lm` command

```
lm(fat ~ protein, data = nnyfs)
```

```
Call:  
lm(formula = fat ~ protein, data = nnyfs)  
  
Coefficients:  
(Intercept)      protein  
    18.8945        0.7251
```

So the fitted line is specified as

$$\text{fat} = 18.8945 + 0.7251 \text{ protein}$$

A detailed summary of the fitted linear regression model is also available.

```
summary(lm(fat ~ protein, data = nnyfs))
```

```
Call:  
lm(formula = fat ~ protein, data = nnyfs)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-77.798 -14.841 -2.449  13.601 110.597  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 18.8945    1.5330   12.32 <2e-16 ***  
protein      0.7251    0.0208   34.87 <2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 25.08 on 1516 degrees of freedom  
Multiple R-squared:  0.4451,    Adjusted R-squared:  0.4447  
F-statistic: 1216 on 1 and 1516 DF,  p-value: < 2.2e-16
```

The way we'll usually summarize the estimated coefficients of a linear model is to use the `broom` package's `tidy` function to put the coefficient estimates into a tibble.

```
tidy(lm(fat ~ protein, data = nnyfs),  
     conf.int = TRUE, conf.level = 0.95) |>  
kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	18.895	1.533	12.325	0	15.887	21.902
protein	0.725	0.021	34.868	0	0.684	0.766

We can also summarize the quality of fit in a linear model using the `broom` package's `glance` function. For now, we'll focus our attention on just one of the many summaries available for a linear model from `glance`: the R-squared value.

```
glance(lm(fat ~ protein, data = nnyfs)) |> select(r.squared) |>  
kable(digits = 3)
```

r.squared
0.445

We'll spend a lot of time working with these regression summaries in this course.

For now, it will suffice to understand the following:

- The outcome variable in this model is `fat`, and the predictor variable is `protein`.
- The straight line model for these data fitted by least squares is $\text{fat} = 18.9 + 0.725 \text{protein}$
- The slope of `protein` is positive, which indicates that as `protein` increases, we expect that `fat` will also increase. Specifically, we expect that for every additional gram of protein consumed, the fat consumption will be 0.725 gram larger.
- The multiple R-squared (squared correlation coefficient) is 0.445, which implies that 44.5% of the variation in `fat` is explained using this linear model with `protein`.
- This also implies that the Pearson correlation between `fat` and `protein` is the square root of 0.445, or 0.667. More on the Pearson correlation soon.

So, if we plan to use a simple (least squares) linear regression model to describe fat consumption as a function of protein consumption in the NNYFS data, does it look like a least squares (or linear regression) model will be an effective choice?

11.3 Correlation Coefficients

Two different correlation measures are worth our immediate attention.

- The one most often used is called the *Pearson* correlation coefficient, and is symbolized with the letter r or sometimes the Greek letter rho (ρ).
- Another tool is the Spearman rank correlation coefficient, also occasionally symbolized by ρ .

For the `nnyfs` data, the Pearson correlation of `fat` and `protein` can be found using the `cor()` function.

```
nnyfs |> select(fat, protein) |> cor()
```

```
      fat    protein
fat    1.0000000 0.6671209
protein 0.6671209 1.0000000
```

Note that the correlation of any variable with itself is 1, and that the correlation of `fat` with `protein` is the same regardless of whether you enter `fat` first or `protein` first.

11.4 The Pearson Correlation Coefficient

Suppose we have n observations on two variables, called X and Y . The Pearson correlation coefficient assesses how well the relationship between X and Y can be described using a linear function.

- The Pearson correlation is **dimension-free**.
- It falls between -1 and +1, with the extremes corresponding to situations where all the points in a scatterplot fall exactly on a straight line with negative and positive slopes, respectively.
- A Pearson correlation of zero corresponds to the situation where there is no linear association.
- Unlike the estimated slope in a regression line, the sample correlation coefficient is symmetric in X and Y , so it does not depend on labeling one of them (Y) the response variable, and one of them (X) the predictor.

Suppose we have n observations on two variables, called X and Y , where \bar{X} is the sample mean of X and s_x is the standard deviation of X . The **Pearson** correlation coefficient r_{XY} is:

$$r_{XY} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

11.5 Studying Correlation through 6 Examples

The `correx1` data file contains six different sets of (x,y) points, identified by the `set` variable.

```
correx1 <- read_csv("data/correx1.csv")
```

```
Rows: 277 Columns: 3
-- Column specification -----
Delimiter: ","
chr (1): set
dbl (2): x, y

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(correx1)
```

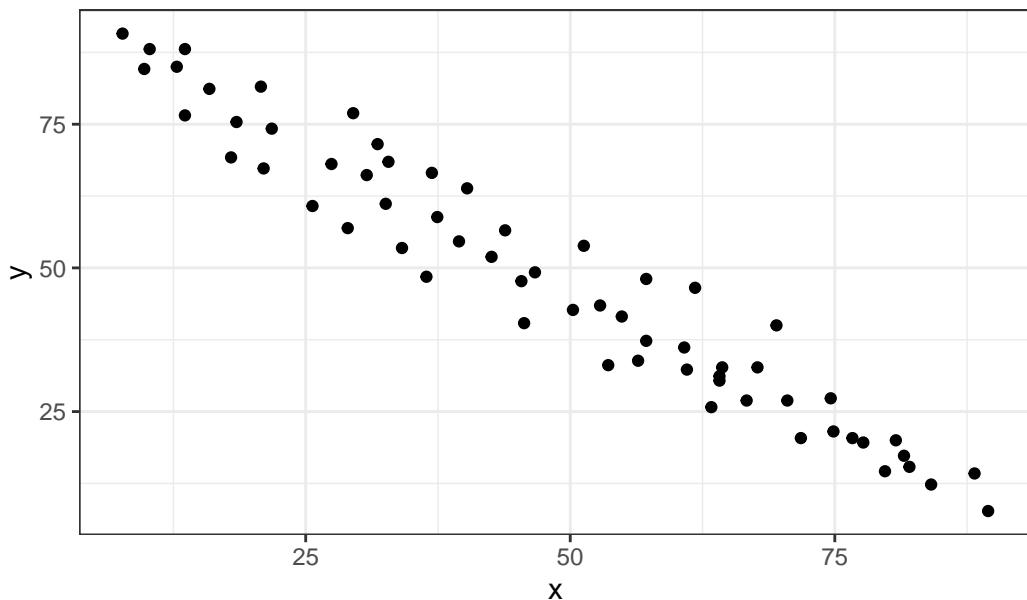
	set	x	y
Length:	277	Min. : 5.897	Min. : 7.308
Class :	character	1st Qu.:29.487	1st Qu.:30.385
Mode :	character	Median :46.154	Median :46.923
		Mean :46.529	Mean :49.061
		3rd Qu.:63.333	3rd Qu.:68.077
		Max. :98.205	Max. :95.385

11.5.1 Data Set Alex

Let's start by working with the **Alex** data set.

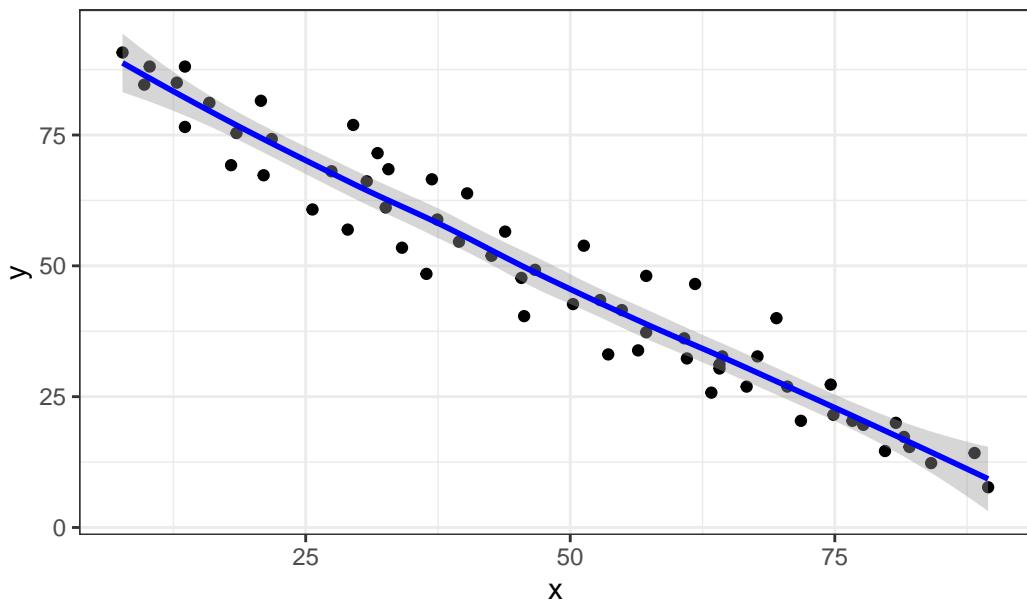
```
ggplot(filter(correx1, set == "Alex"), aes(x = x, y = y)) +
  geom_point() +
  labs(title = "correx1: Data Set Alex")
```

correx1: Data Set Alex



```
ggplot(filter(correx1, set == "Alex"), aes(x = x, y = y)) +  
  geom_point() +  
  geom_smooth(method = "loess", formula = y ~ x, col = "blue") +  
  labs(title = "correx1: Alex, with loess smooth")
```

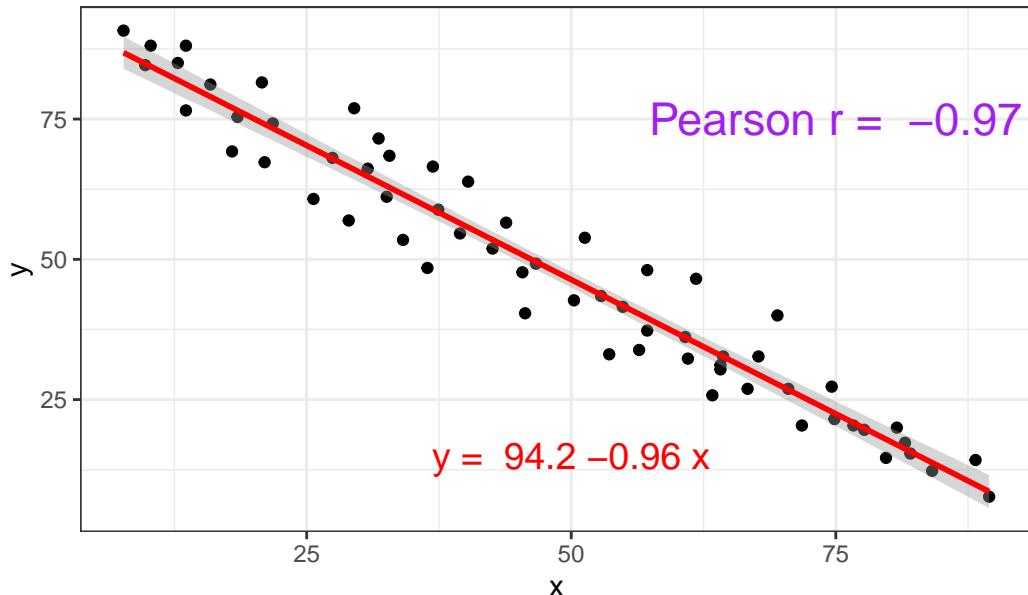
correx1: Alex, with loess smooth



```
setA <- filter(correx1, set == "Alex")

ggplot(setA, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, col = "red") +
  labs(title = "correx1: Alex, with Fitted Linear Model") +
  annotate("text", x = 75, y = 75, col = "purple", size = 6,
          label = paste("Pearson r = ", signif(cor(setA$x, setA$y),3))) +
  annotate("text", x = 50, y = 15, col = "red", size = 5,
          label = paste("y = ", signif(coef(lm(setA$y ~ setA$x))[1],3),
                        signif(coef(lm(setA$y ~ setA$x))[2],2), "x"))
```

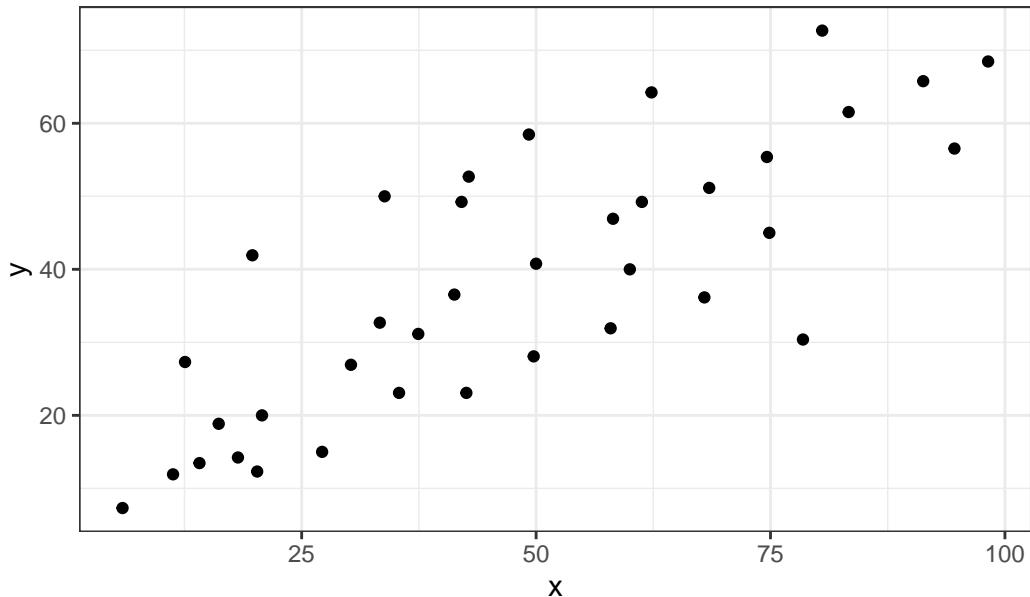
correx1: Alex, with Fitted Linear Model



11.5.2 Data Set Bonnie

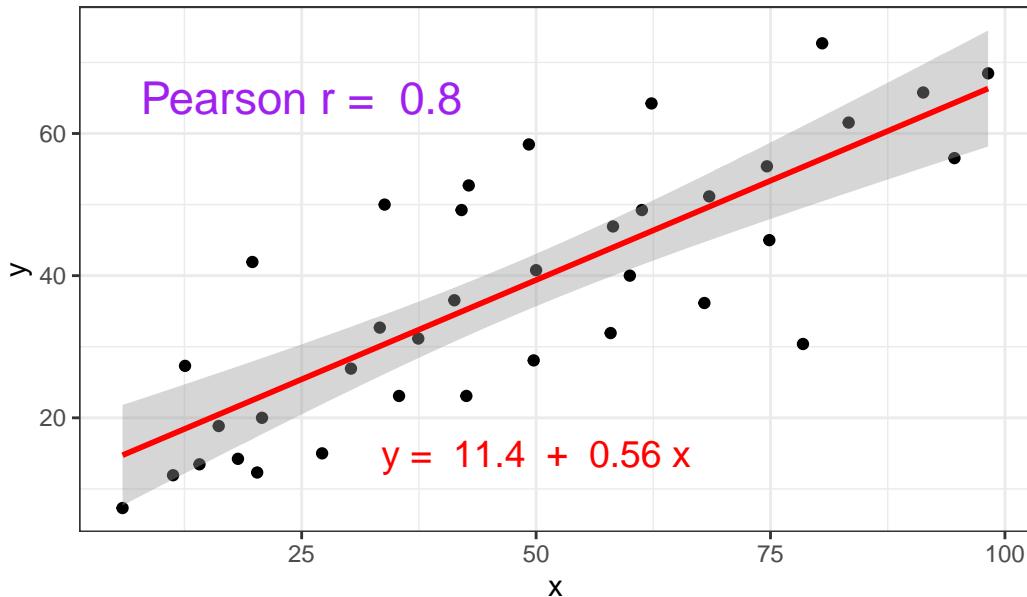
```
setB <- dplyr::filter(correx1, set == "Bonnie")  
  
ggplot(setB, aes(x = x, y = y)) +  
  geom_point() +  
  labs(title = "correx1: Data Set Bonnie")
```

correx1: Data Set Bonnie



```
ggplot(setB, aes(x = x, y = y)) +  
  geom_point() +  
  geom_smooth(method = "lm", formula = y ~ x, col = "red") +  
  labs(title = "correx1: Bonnie, with Fitted Linear Model") +  
  annotate("text", x = 25, y = 65, col = "purple", size = 6,  
          label = paste("Pearson r = ", signif(cor(setB$x, setB$y), 2))) +  
  annotate("text", x = 50, y = 15, col = "red", size = 5,  
          label = paste("y = ", signif(coef(lm(setB$y ~ setB$x))[1], 3),  
                        " + ",  
                        signif(coef(lm(setB$y ~ setB$x))[2], 2), "x"))
```

correx1: Bonnie, with Fitted Linear Model



11.5.3 Correlations for All Six Data Sets in the Correx1 Example

Let's look at the Pearson correlations associated with each of the six data sets contained in the `correx1` example.

```
tab1 <- correx1 |>
  group_by(set) |>
  summarise("Pearson r" = round(cor(x, y, use="complete"), 2))

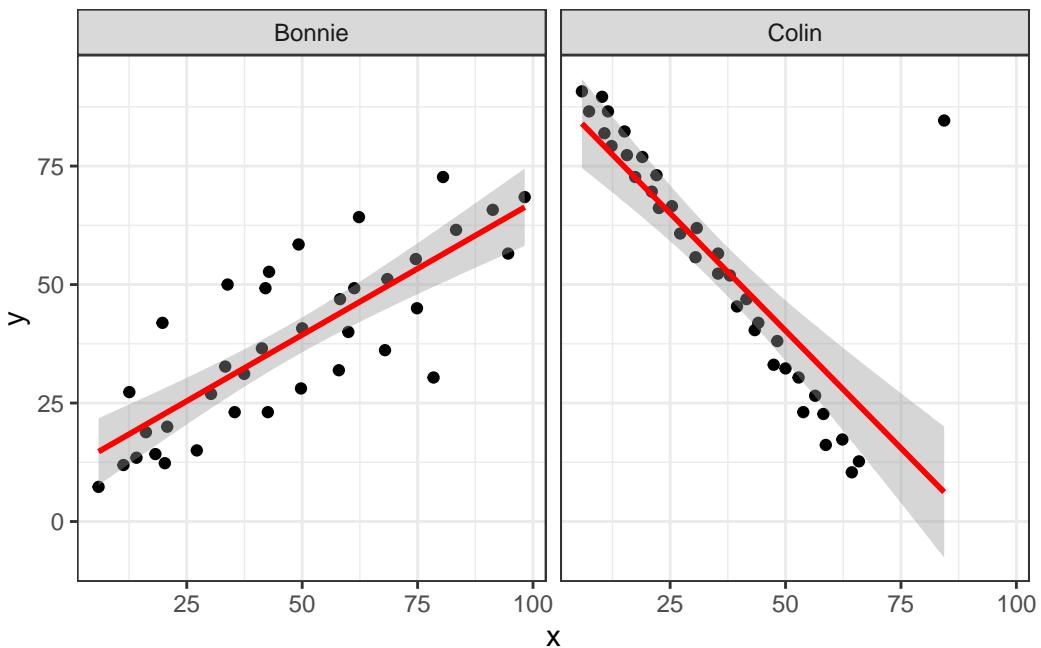
knitr::kable(tab1)
```

set	Pearson r
Alex	-0.97
Bonnie	0.80
Colin	-0.80
Danielle	0.00
Earl	-0.01
Fiona	0.00

11.5.4 Data Set Colin

It looks like the picture for Colin should be very similar (in terms of scatter) to the picture for Bonnie, except that Colin will have a negative slope, rather than the positive one Bonnie has. Is that how this plays out?

```
setBC <- filter(correx1, set == "Bonnie" | set == "Colin")  
  
ggplot(setBC, aes(x = x, y = y)) +  
  geom_point() +  
  geom_smooth(method = "lm", formula = y ~ x, col = "red") +  
  facet_wrap(~ set)
```

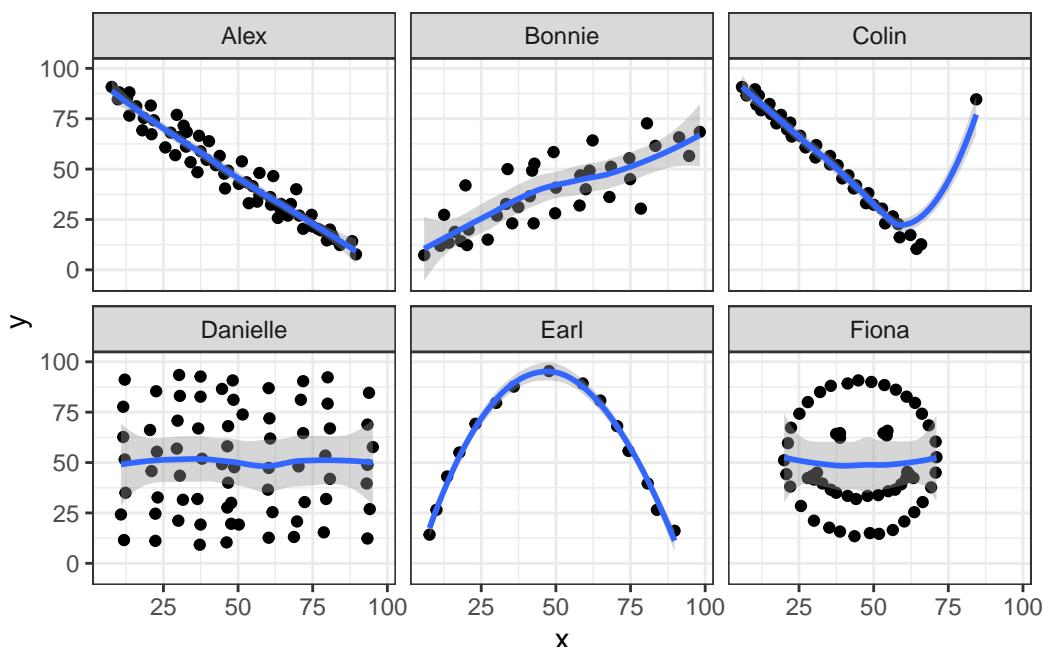


Uh, oh. It looks like the point in Colin at the top right is twisting what would otherwise be a very straight regression model with an extremely strong negative correlation. There's no better way to look for outliers than to examine the scatterplot.

11.5.5 Draw the Picture!

We've seen that Danielle, Earl and Fiona all show Pearson correlations of essentially zero. However, the three data sets look very different in a scatterplot.

```
ggplot(correx1, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "loess", formula = y ~ x) +
  facet_wrap(~ set)
```



When we learn that the correlation is zero, we tend to assume we have a picture like the Danielle data set. If Danielle were our real data, we might well think that x would be of little use in predicting y .

- But what if our data looked like Earl? In the Earl data set, x is incredibly helpful in predicting y , but we can't use a straight line model - instead, we need a non-linear modeling approach.
- You'll recall that the Fiona data set also had a Pearson correlation of zero. But here, the picture is rather more interesting.

So, remember, draw the appropriate scatterplot whenever you make use of a correlation coefficient.

```
rm(setA, setB, setBC, tab1)
```

11.6 Estimating Correlation from Scatterplots

The correx2 data set is designed to help you calibrate yourself a bit in terms of estimating a correlation from a scatterplot. There are 11 data sets buried within the correx2 example, and they are labeled by their Pearson correlation coefficients, ranging from $r = 0.01$ to $r = 0.999$

```
correx2 <- read_csv("data/correx2.csv")
```

Rows: 582 Columns: 4

-- Column specification -----

Delimiter: ","

chr (1): set

dbl (3): x, y, group

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```
correx2 |>
  group_by(set) |>
  summarise(cor = round(cor(x, y, use="complete"), 3))
```

A tibble: 11 x 2

set cor

<chr> <dbl>

1 Set 01 0.01

2 Set 10 0.102

3 Set 20 0.202

4 Set 30 0.301

5 Set 40 0.403

6 Set 50 0.499

7 Set 60 0.603

8 Set 70 0.702

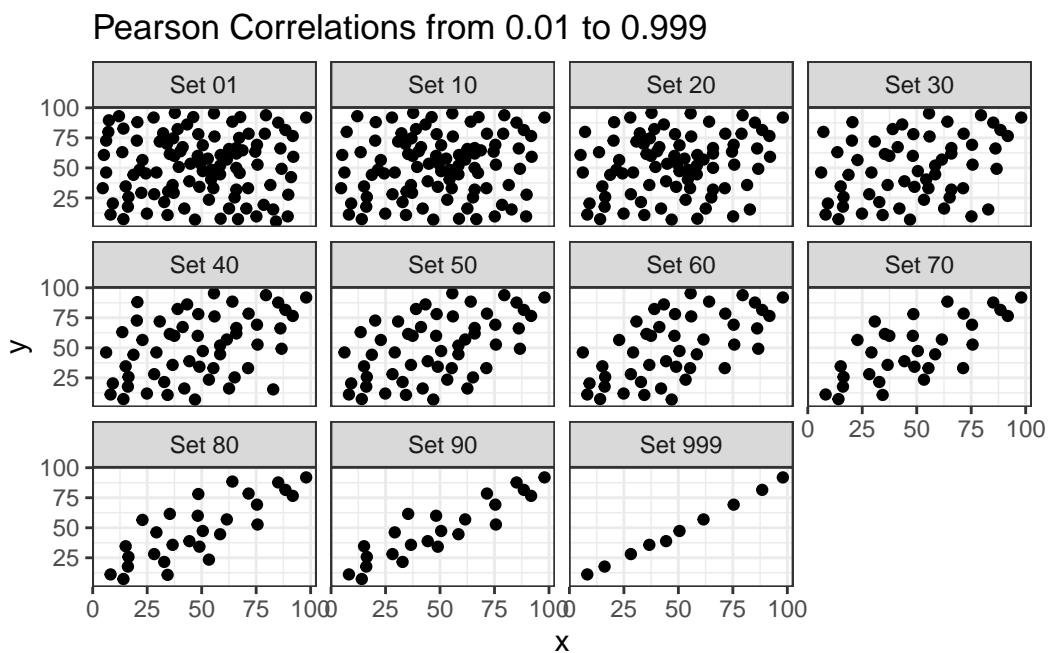
9 Set 80 0.799

10 Set 90 0.902

11 Set 999 0.999

Here is a plot of the 11 data sets, showing the increase in correlation from 0.01 (in Set 01) to 0.999 (in Set 999).

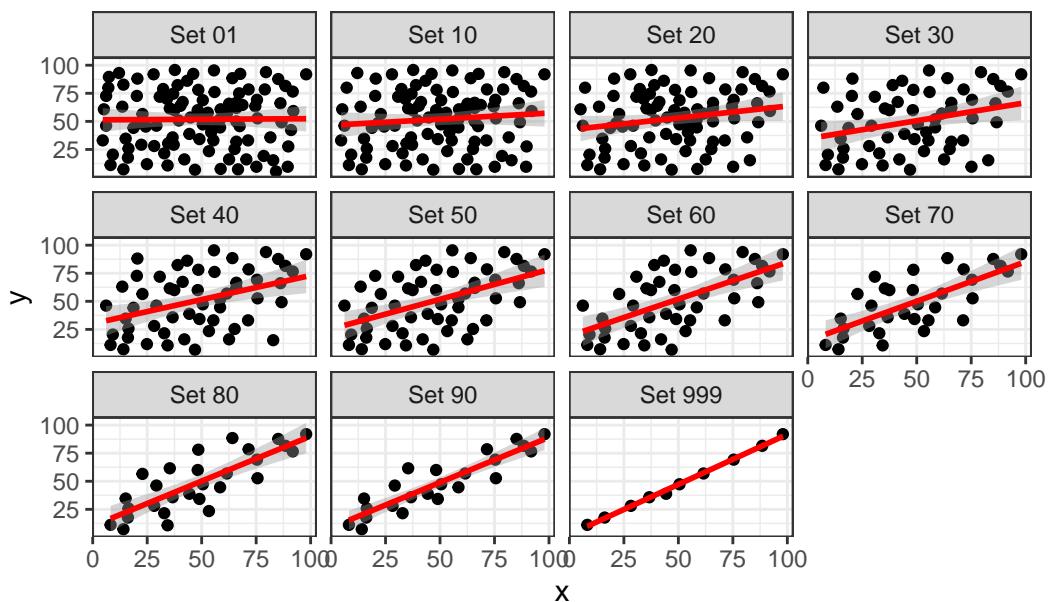
```
ggplot(correx2, aes(x = x, y = y)) +
  geom_point() +
  facet_wrap(~ set) +
  labs(title = "Pearson Correlations from 0.01 to 0.999")
```



Note that R will allow you to fit a straight line model to any of these relationships, no matter how appropriate it might be to do so.

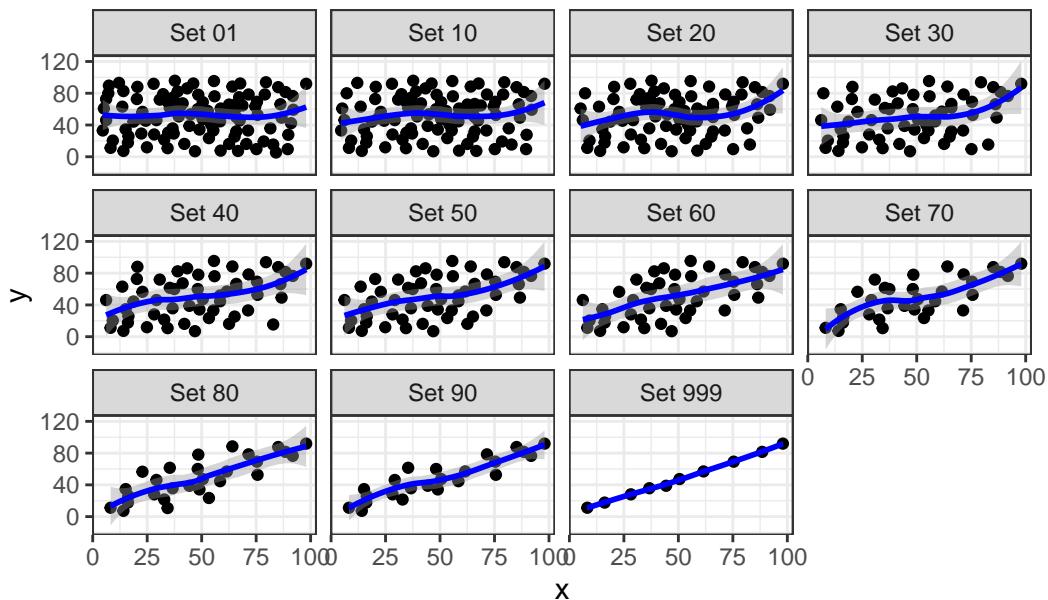
```
ggplot(correx2, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, col = "red") +
  facet_wrap(~ set) +
  labs(title = "R will fit a straight line to anything.")
```

R will fit a straight line to anything.



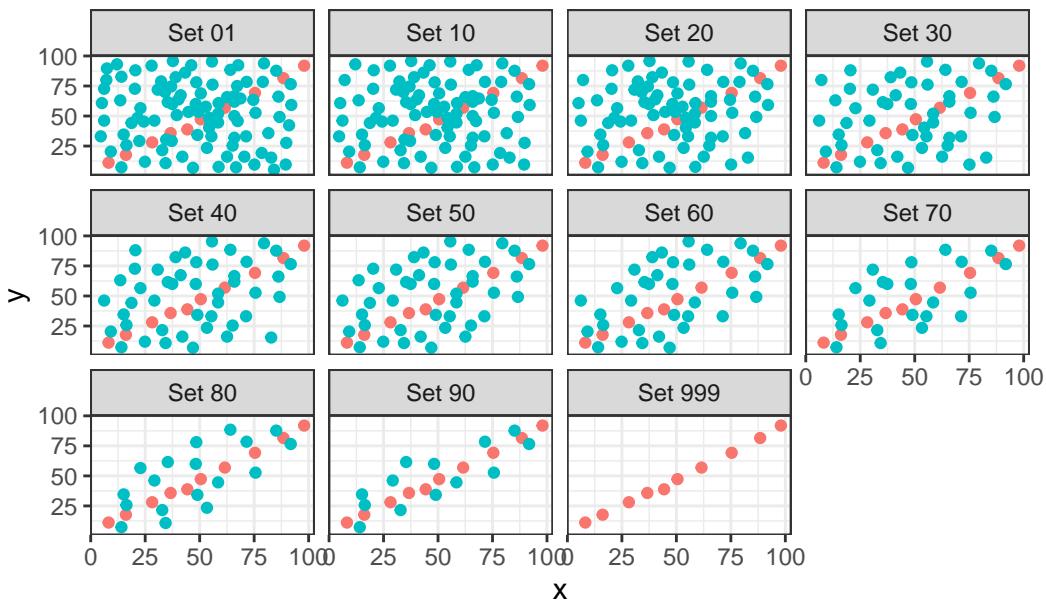
```
ggplot(correx2, aes(x = x, y = y)) +  
  geom_point() +  
  geom_smooth(col = "blue") +  
  facet_wrap(~ set) +  
  labs(title = "Even if a loess smooth suggests non-linearity.")
```

Even if a loess smooth suggests non-linearity.



```
ggplot(correx2, aes(x = x, y = y, color = factor(group))) +  
  geom_point() +  
  guides(color = "none") +  
  facet_wrap(~ set) +  
  labs(title = "Note: The same 10 points (in red) are in each plot.")
```

Note: The same 10 points (in red) are in each plot.



Note that the same 10 points are used in each of the data sets. It's always possible that a lurking subgroup of the data within a scatterplot follows a very strong linear relationship. This is why it's so important (and difficult) not to go searching for such a thing without a strong foundation of logic, theory and prior empirical evidence.

11.7 The Spearman Rank Correlation

The Spearman rank correlation coefficient is a rank-based measure of statistical dependence that assesses how well the relationship between X and Y can be described using a **monotone function** even if that relationship is not linear.

- A monotone function preserves order, that is, Y must either be strictly increasing as X increases, or strictly decreasing as X increases.
- A Spearman correlation of 1.0 indicates simply that as X increases, Y always increases.
- Like the Pearson correlation, the Spearman correlation is dimension-free, and falls between -1 and +1.
- A positive Spearman correlation corresponds to an increasing (but not necessarily linear) association between X and Y, while a negative Spearman correlation corresponds to a decreasing (but again not necessarily linear) association.

11.7.1 Spearman Formula

To calculate the Spearman rank correlation, we take the ranks of the X and Y data, and then apply the usual Pearson correlation. To find the ranks, sort X and Y into ascending order, and then number them from 1 (smallest) to n (largest). In the event of a tie, assign the average rank to the tied subjects.

11.7.2 Comparing Pearson and Spearman Correlations

Let's look at the `nnyfs` data again.

```
nnyfs |> select(fat, protein) |> cor()

      fat    protein
fat    1.0000000 0.6671209
protein 0.6671209 1.0000000

nnyfs |> select(fat, protein) %>% cor(., method = "spearman")

      fat    protein
fat    1.0000000 0.6577489
protein 0.6577489 1.0000000
```

The Spearman and Pearson correlations are not especially different in this case.

11.7.3 Spearman vs. Pearson Example 1

The next few plots describe relationships where we anticipate the Pearson and Spearman correlations might differ in their conclusions.

```
spear1 <- read_csv("data/spear1.csv")

Rows: 22 Columns: 2
-- Column specification -----
Delimiter: ","
dbl (2): x, y

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
spear2 <- read_csv("data/spear2.csv")
```

Rows: 90 Columns: 2

-- Column specification -----

Delimiter: ","

dbl (2): x, y

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```
spear3 <- read_csv("data/spear3.csv")
```

Rows: 55 Columns: 2

-- Column specification -----

Delimiter: ","

dbl (2): x, y

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```
spear4 <- read_csv("data/spear4.csv")
```

Rows: 15 Columns: 2

-- Column specification -----

Delimiter: ","

dbl (2): x, y

i Use `spec()` to retrieve the full column specification for this data.

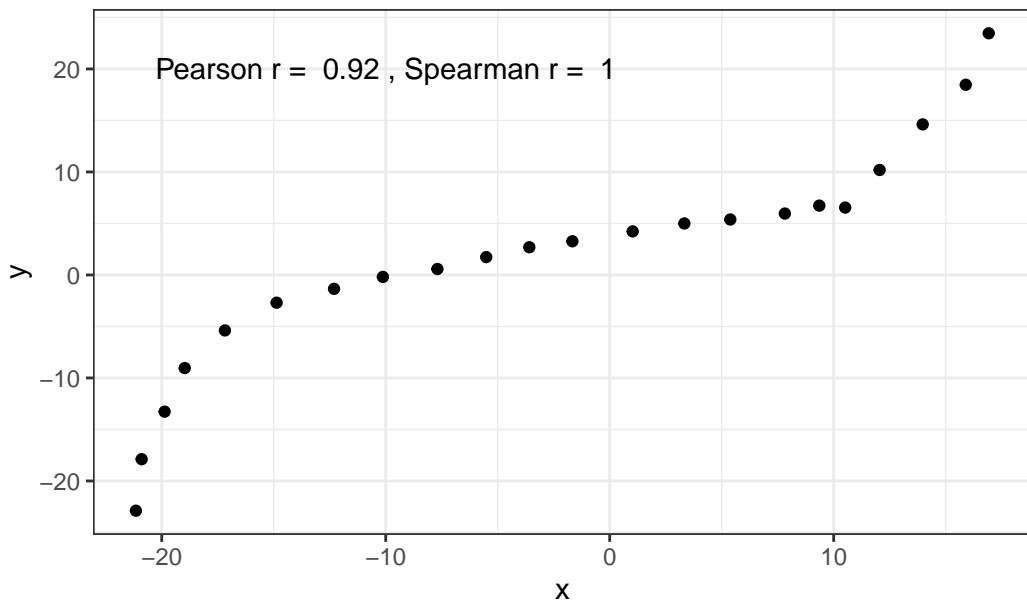
i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```
# these are just toy examples with  
# two columns per data set and no row numbering
```

Example 1 shows a function where the Pearson correlation is 0.925 (a strong but not perfect linear relation), but the Spearman correlation is 1 because the relationship is monotone, even though it is not perfectly linear.

```
ggplot(spear1, aes(x = x, y = y)) +
  geom_point() +
  labs(title = "Spearman vs. Pearson, Example 1") +
  annotate("text", x = -10, y = 20,
           label = paste("Pearson r = ",
                         signif(cor(spear1$x, spear1$y), 2),
                         ", Spearman r = ",
                         signif(cor(spear1$x, spear1$y, method = "spearman"), 2)))
```

Spearman vs. Pearson, Example 1



So, a positive Spearman correlation corresponds to an increasing (but not necessarily linear) association between x and y.

11.7.4 Spearman vs. Pearson Example 2

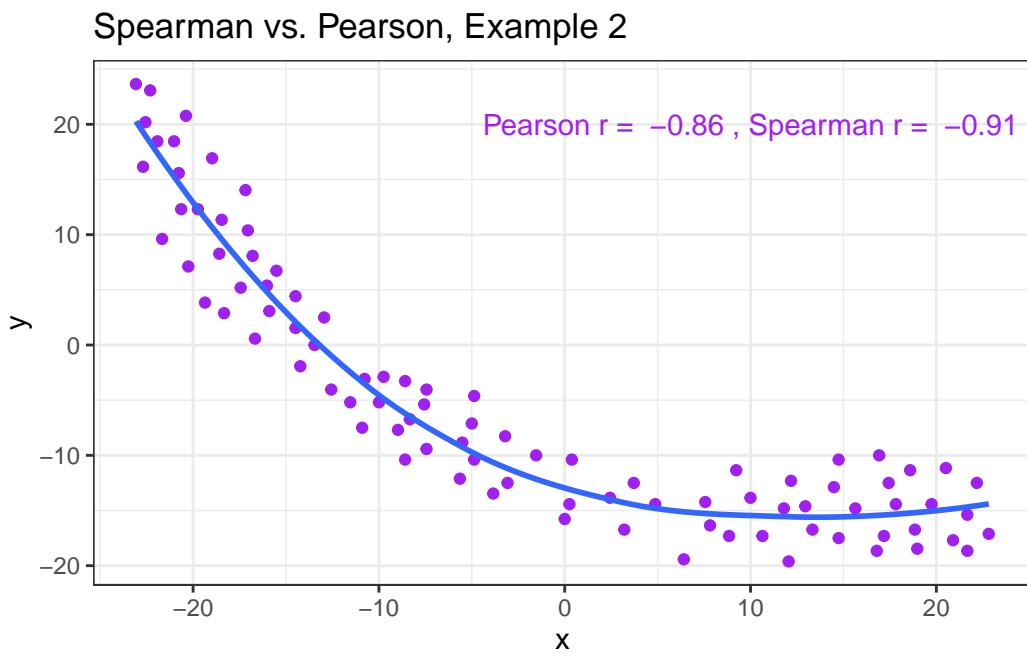
Example 2 shows that a negative Spearman correlation corresponds to a decreasing (but, again, not necessarily linear) association between x and y.

```
ggplot(spear2, aes(x = x, y = y)) +
  geom_point(col = "purple") +
  geom_smooth(method = "loess", formula = y ~ x, se = FALSE) +
```

```

labs(title = "Spearman vs. Pearson, Example 2") +
annotate("text", x = 10, y = 20, col = "purple",
label = paste("Pearson r = ",
signif(cor(spear2$x, spear2$y),2),
", Spearman r = ",
signif(cor(spear2$x, spear2$y, method = "spearman"),2)))

```



11.7.5 Spearman vs. Pearson Example 3

The Spearman correlation is less sensitive than the Pearson correlation is to strong outliers that are unusual on either the X or Y axis, or both. That is because the Spearman rank coefficient limits the outlier to the value of its rank.

In Example 3, for instance, the Spearman correlation reacts much less to the outliers around $X = 12$ than does the Pearson correlation.

```

ggplot(spear3, aes(x = x, y = y)) +
  geom_point(col = "blue") +
  labs(title = "Spearman vs. Pearson, Example 3") +
  annotate("text", x = 5, y = -15, col = "blue",
  label = paste("Pearson r = ",

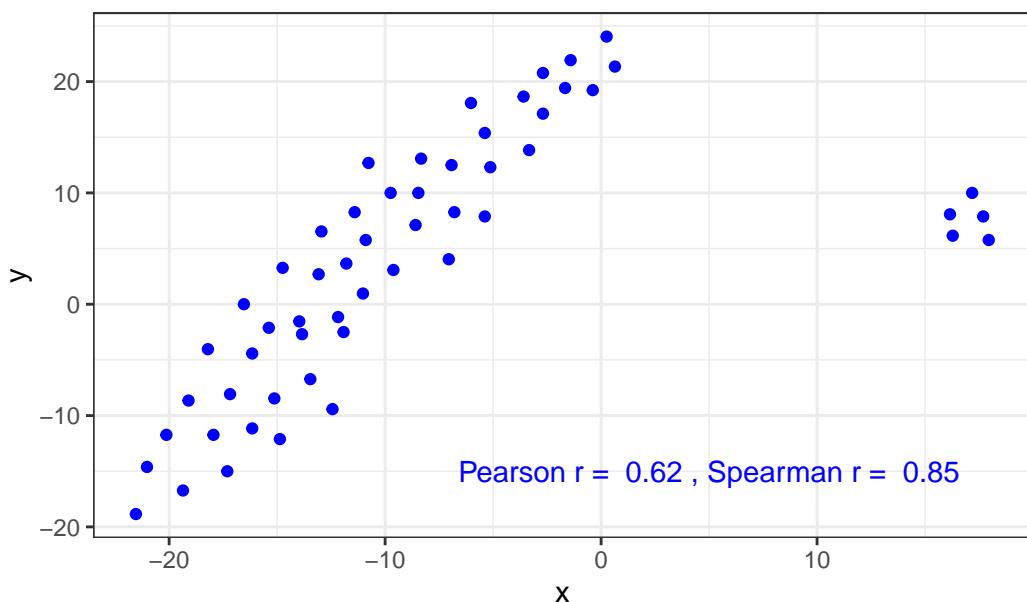
```

```

signif(cor(spear3$x, spear3$y), 2),
", Spearman r = ",
signif(cor(spear3$x, spear3$y, method = "spearman"), 2)))

```

Spearman vs. Pearson, Example 3



11.7.6 Spearman vs. Pearson Example 4

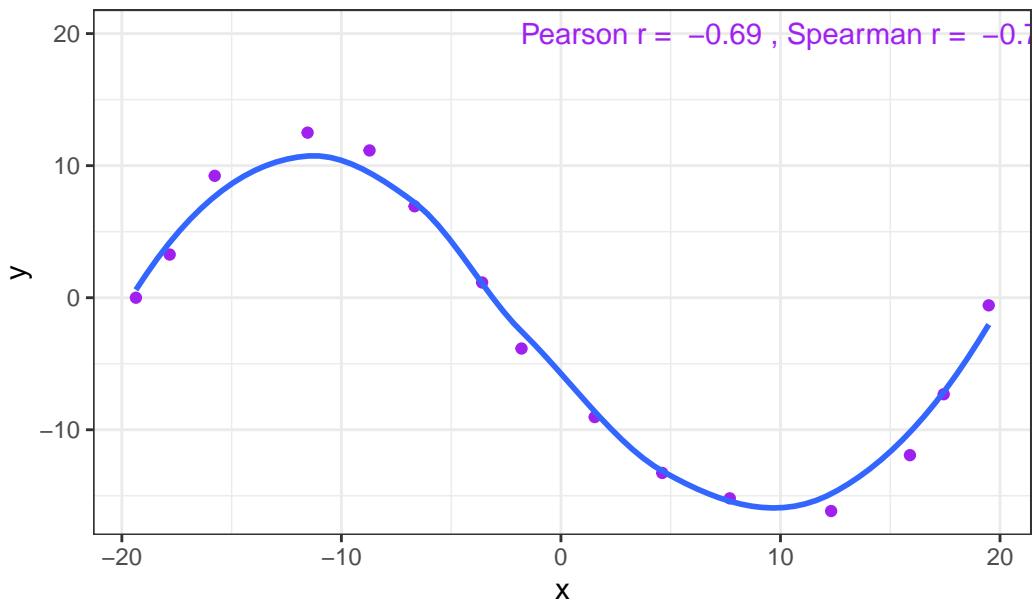
The use of a Spearman correlation is no substitute for looking at the data. For non-monotone data like what we see in Example 4, neither the Spearman nor the Pearson correlation alone provides much guidance, and just because they are (essentially) telling you the same thing, that doesn't mean what they're telling you is all that helpful.

```

ggplot(spear4, aes(x = x, y = y)) +
  geom_point(col = "purple") +
  geom_smooth(method = "loess", formula = y ~ x, se = FALSE) +
  labs(title = "Spearman vs. Pearson, Example 4") +
  annotate("text", x = 10, y = 20, col = "purple",
    label = paste("Pearson r = ",
      signif(cor(spear4$x, spear4$y), 2),
      ", Spearman r = ",
      signif(cor(spear4$x, spear4$y, method = "spearman"), 2)))

```

Spearman vs. Pearson, Example 4



12 Linearizing Transformations

12.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(broom)
library(car)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

12.2 “Linearize” The Association between Quantitative Variables

Confronted with a scatterplot describing a monotone association between two quantitative variables, we may decide the data are not well approximated by a straight line, and thus, that a least squares regression may not be sufficiently useful. In these circumstances, we have at least two options, which are not mutually exclusive:

- a. Let the data be as they may, and summarize the scatterplot using tools like loess curves, polynomial functions, or cubic splines to model the relationship.
- b. Consider re-expressing the data (often we start with re-expressions of the outcome data [the Y variable]) using a transformation so that the transformed data may be modeled effectively using a straight line.

12.3 The Box-Cox Plot

As before, Tukey’s ladder of power transformations can guide our exploration.

Power (λ)	-2	-1	-1/2	0	1/2	1	2
Transformation	$1/y^2$	$1/y$	$1/\sqrt{y}$	$\log y$	\sqrt{y}	y	y^2

The **Box-Cox plot**, from the `boxCox` function in the `car` package, sifts through the ladder of options to suggest a transformation (for Y) to best linearize the outcome-predictor(s) relationship.

12.3.1 A Few Caveats

1. These methods work well with *monotone* data, where a smooth function of Y is either strictly increasing, or strictly decreasing, as X increases.
2. Some of these transformations require the data to be positive. We can rescale the Y data by adding a constant to every observation in a data set without changing shape.
3. We can use a natural logarithm (`log` in R), a base 10 logarithm (`log10`) or even sometimes a base 2 logarithm (`log2`) to good effect in Tukey's ladder. All affect the association's shape in the same way, so we'll stick with `log` (base e).
4. Some re-expressions don't lead to easily interpretable results. Not many things that make sense in their original units also make sense in inverse square roots. There are times when we won't care, but often, we will.
5. If our primary interest is in making predictions, we'll generally be more interested in getting good predictions back on the original scale, and we can back-transform the point and interval estimates to accomplish this.

12.4 A Simulated Example

```

set.seed(999);
x.rand <- rbeta(80, 2, 5) * 20 + 3
set.seed(1000);
y.rand <- abs(50 + 0.75*x.rand^(3)
              - 0.65*x.rand + rnorm(80, 0, 200))

scatter1 <- tibble(x = x.rand, y = y.rand)
rm(x.rand, y.rand)

ggplot(scatter1, aes(x = x, y = y)) +
  geom_point(shape = 1, size = 3) +
  ## add loess smooth

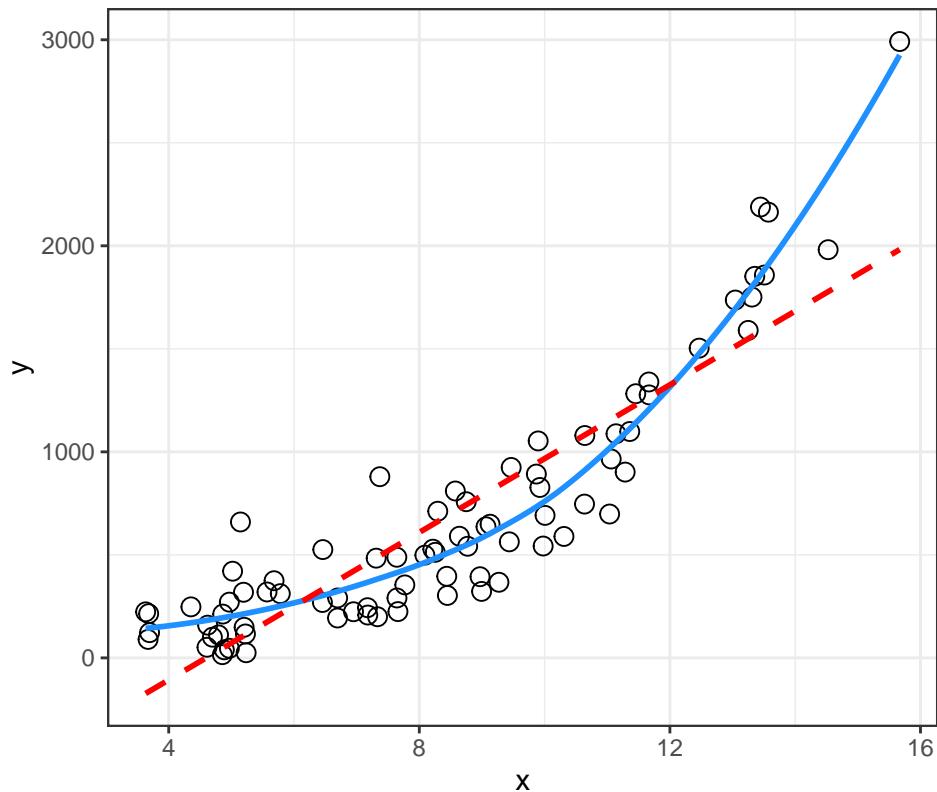
```

```

geom_smooth(method = "loess", se = FALSE,
            col = "dodgerblue", formula = y ~ x) +
## then add linear fit
geom_smooth(method = "lm", se = FALSE,
            col = "red", formula = y ~ x, linetype = "dashed") +
labs(title = "Simulated scatter1 example: Y vs. X")

```

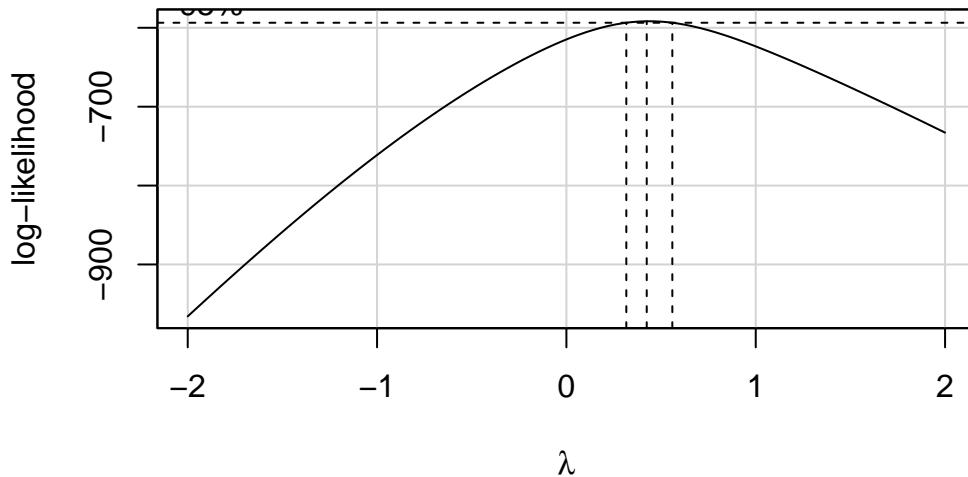
Simulated scatter1 example: Y vs. X



Having simulated data that produces a curved scatterplot, I will now use the Box-Cox plot to lead my choice of an appropriate power transformation for Y in order to “linearize” the association of Y and X.

```
boxCox(scatter1$y ~ scatter1$x)
```

Profile Log-likelihood



```
powerTransform(scatter1$y ~ scatter1$x)
```

```
Estimated transformation parameter
Y1
0.4368753
```

The Box-Cox plot peaks at the value $\lambda = 0.44$, which is pretty close to $\lambda = 0.5$. Now, 0.44 isn't on Tukey's ladder, but 0.5 is.

Power (λ)	-2	-1	-1/2	0	1/2	1	2
Transformation/y ²	1/y	1/ \sqrt{y}	log y	\sqrt{y}	y	y^2	

If we use $\lambda = 0.5$, on Tukey's ladder of power transformations, it suggests we look at the relationship between the square root of Y and X, as shown next.

```
p1 <- ggplot(scatter1, aes(x = x, y = y)) +
  geom_point(size = 2) +
  geom_smooth(method = "loess", se = FALSE,
              formula = y ~ x, col = "dodgerblue") +
```

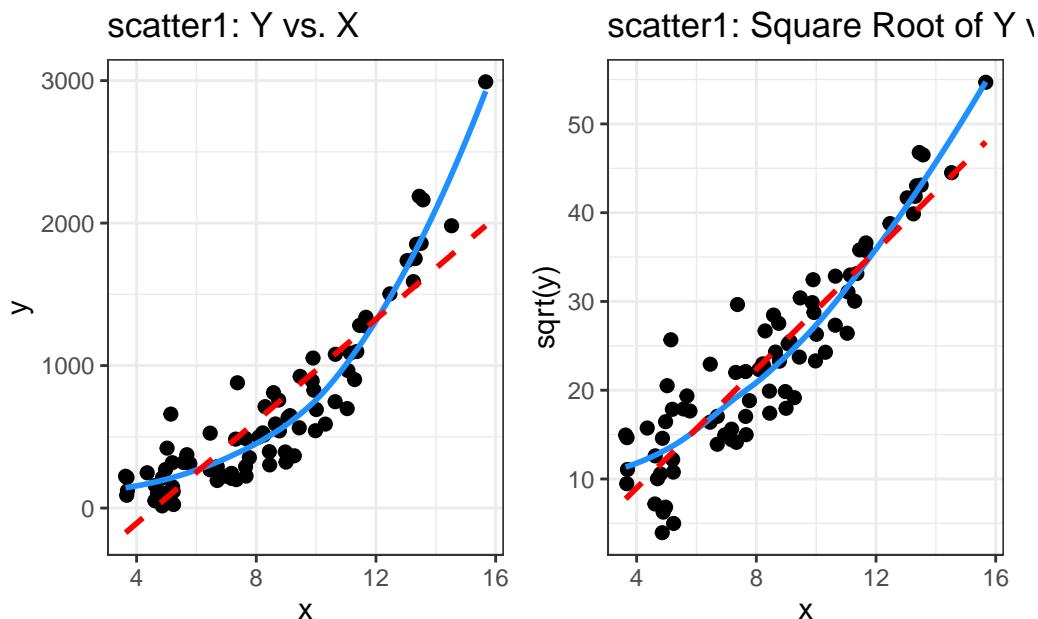
```

geom_smooth(method = "lm", se = FALSE,
            formula = y ~ x, col = "red", linetype = "dashed") +
  labs(title = "scatter1: Y vs. X")

p2 <- ggplot(scatter1, aes(x = x, y = sqrt(y))) +
  geom_point(size = 2) +
  geom_smooth(method = "loess", se = FALSE,
              formula = y ~ x, col = "dodgerblue") +
  geom_smooth(method = "lm", se = FALSE,
              formula = y ~ x, col = "red", linetype = "dashed") +
  labs(title = "scatter1: Square Root of Y vs. X")

p1 + p2

```



By eye, I think the square root plot better matches the linear fit.

12.5 Checking on a Transformation or Re-Expression

We can do three more things to check on our transformation.

1. We can calculate the correlation of our original and re-expressed associations.

2. We can use the `testTransform` function in the `car` library in R to perform a statistical test comparing the optimal choice of power ($\lambda = 0.44$) to various other transformations.
3. We can go ahead and fit the regression models using each approach and compare the plots of studentized residuals against fitted values from the data to see if the re-expression reduces the curve in that residual plot, as well.

Option 3 is by far the most important in practice, and it's the one we'll focus on going forward, but we'll demonstrate all three here.

12.5.1 Checking the Correlation Coefficients

Here, we calculate the correlation of original and re-expressed associations.

```
cor(scatter1$y, scatter1$x)

[1] 0.891198

cor(sqrt(scatter1$y), scatter1$x)

[1] 0.9144307
```

The Pearson correlation is a little stronger after the transformation. as we'd expect.

12.5.2 Using the `testTransform` function

Here, we use the `testTransform` function (also from the `car` package) to compare the optimal choice determined by the `powerTransform` function (here $\lambda = 0.44$) to $\lambda = 0$ (logarithm), 0.5 (square root) and 1 (no transformation).

```
testTransform(powerTransform(scatter1$y ~ scatter1$x), 0)

      LRT df      pval
LR test, lambda = (0) 46.17947 1 1.079e-11

testTransform(powerTransform(scatter1$y ~ scatter1$x), 0.5)

      LRT df      pval
LR test, lambda = (0.5) 1.024888 1 0.31136
```

```
testTransform(powerTransform(scatter1$y ~ scatter1$x), 1)
```

```
          LRT df      pval
LR test, lambda = (1) 63.75953 1 1.4433e-15
```

- It looks like only the square root ($\lambda = 0.5$) of these three options is not significantly worse by the log-likelihood criterion applied here than the optimal choice.
- That's because it's the only one with a p value larger than our usual standard for statistical significance, of 0.05.

12.5.3 Comparing the Residual Plots

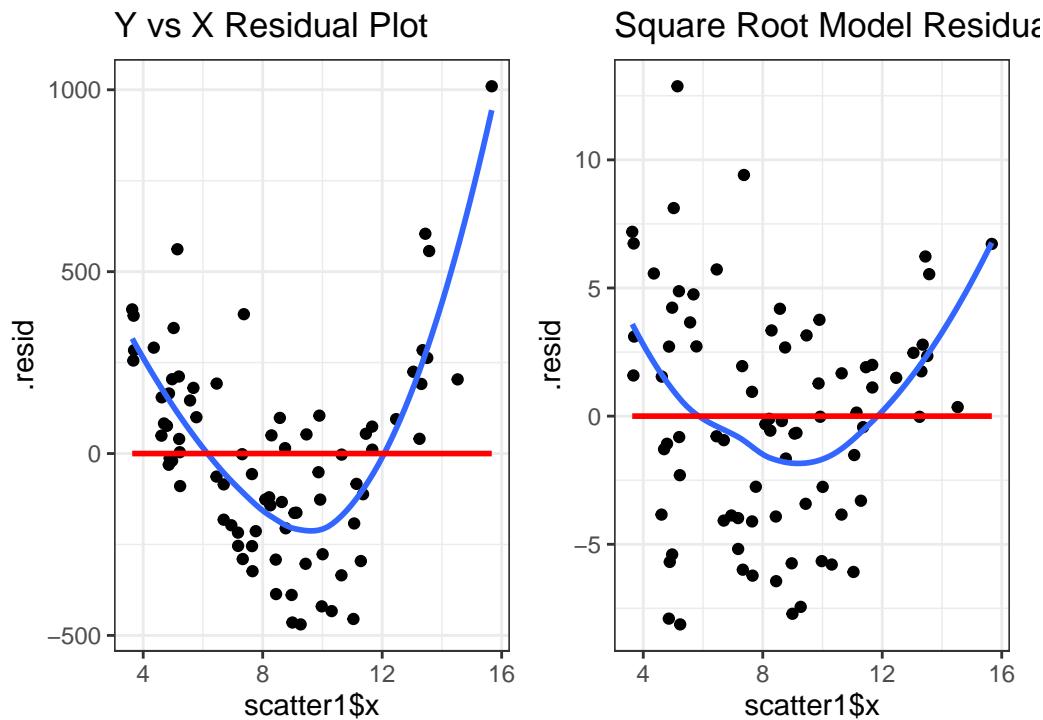
We can fit the regression models, obtain plots of residuals against fitted values, and compare them to see which one has less indication of a curve in the residuals.

```
model.orig <- lm(scatter1$y ~ scatter1$x)
model.sqrt <- lm(sqrt(scatter1$y) ~ scatter1$x)

p1 <- augment(model.orig) %>%
  ggplot(., aes(x = scatter1$x, y = .resid)) +
  geom_point() +
  geom_smooth(method = "loess", formula = y ~ x, se = FALSE) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, col = "red") +
  labs(title = "Y vs X Residual Plot")

p2 <- augment(model.sqrt) %>%
  ggplot(., aes(x = scatter1$x, y = .resid)) +
  geom_point() +
  geom_smooth(method = "loess", formula = y ~ x, se = FALSE) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, col = "red") +
  labs(title = "Square Root Model Residuals")

p1 + p2
```



What we're looking for in such a plot is the absence of a curve, among other things, we want to see “fuzzy football” shapes.

As compared to the original residual plot, the square root version, is a modest improvement in this regard. It does look a bit less curved, and a bit more like a random cluster of points, so that's nice. Usually, we can do a little better in real data, as shown in the next example from the NNYFS data we introduced in Chapter @ref(NYFS-Study).

12.6 An Example from the NNYFS data

```
nnyfs <- read_rds("data/nnyfs.Rds")
```

Using the subjects in the `nnyfs` data with complete data on the two variables of interest, let's look at the relationship between arm circumference (the outcome, shown on the Y axis) and arm length (the predictor, shown on the X axis.)

```
nnyfs_c <- nnyfs |>  
  filter(complete.cases(arm_circ, arm_length)) |>  
  select(SEQN, arm_circ, arm_length)
```

12.6.1 Pearson correlation and scatterplot

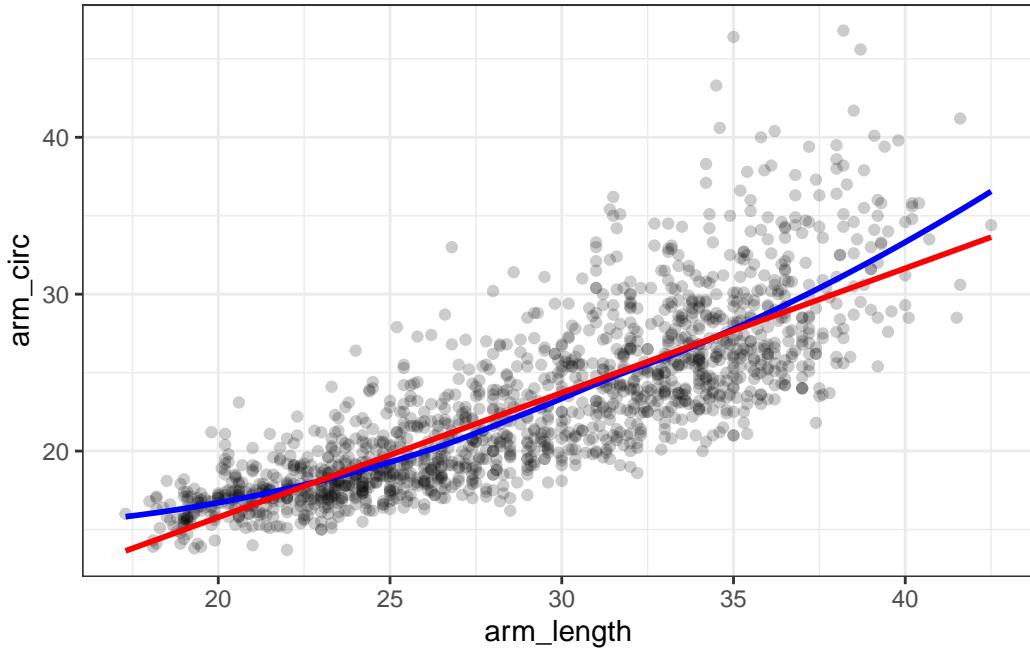
Here is the Pearson correlation between these two variables.

```
nnyfs_c |> select(arm_length, arm_circ) |> cor()
```

	arm_length	arm_circ
arm_length	1.0000000	0.8120242
arm_circ	0.8120242	1.0000000

Here's the resulting scatterplot.

```
ggplot(nnyfs_c, aes(x = arm_length, y = arm_circ)) +  
  geom_point(alpha = 0.2) +  
  geom_smooth(method = "loess", formula = y ~ x,  
              se = FALSE, color = "blue") +  
  geom_smooth(method = "lm", formula = y ~ x,  
              se = FALSE, color = "red")
```



While the Pearson correlation is still quite strong, note that the loess smooth (shown in blue) bends up from the straight line model (shown in red) at both the low and high end of arm length.

Note also the use of `alpha = 0.2` to show the points with greater transparency than they would be shown normally (the default setting is no transparency with `alpha = 1.`)

12.6.2 Plotting the Residuals

Now, let's build a plot of residuals from the straight line model plotted against the arm length. We can obtain these residuals using the `augment()` function from the `broom` package.

```
m1 <- lm(arm_circ ~ arm_length, data = nnyfs_c)

nnyfs_c_aug1 <- augment(m1, data = nnyfs_c)

nnyfs_c_aug1

# A tibble: 1,511 x 9
  SEQN arm_circ arm_length .fitted .resid      .hat .sigma   .cooksdi .std.resid
  <dbl>    <dbl>     <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
```

```

1 71918    25.4      27.7    21.9  3.51  0.000695  3.21 0.000416   1.09
2 71919    26        38.4    30.4 -4.38  0.00253   3.21 0.00237  -1.37
3 71920    37.9      35.9    28.4  9.50  0.00167   3.20 0.00735   2.96
4 71921    15.1      18.3    14.4  0.669 0.00304   3.21 0.0000663  0.209
5 71922    29.5      34.2    27.0  2.45  0.00124   3.21 0.000362  0.764
6 71923    27.9      33       26.1  1.80  0.00100   3.21 0.000159  0.562
7 71924    17.6      26.5    20.9 -3.34  0.000788  3.21 0.000427  -1.04
8 71925    17.7      24.2    19.1 -1.41  0.00113   3.21 0.000110  -0.441
9 71926    19.9      26       20.5 -0.642 0.000844  3.21 0.0000169  -0.200
10 71927   17.3      20       15.8  1.52  0.00234   3.21 0.000263  0.474
# ... with 1,501 more rows
# i Use `print(n = ...)` to see more rows

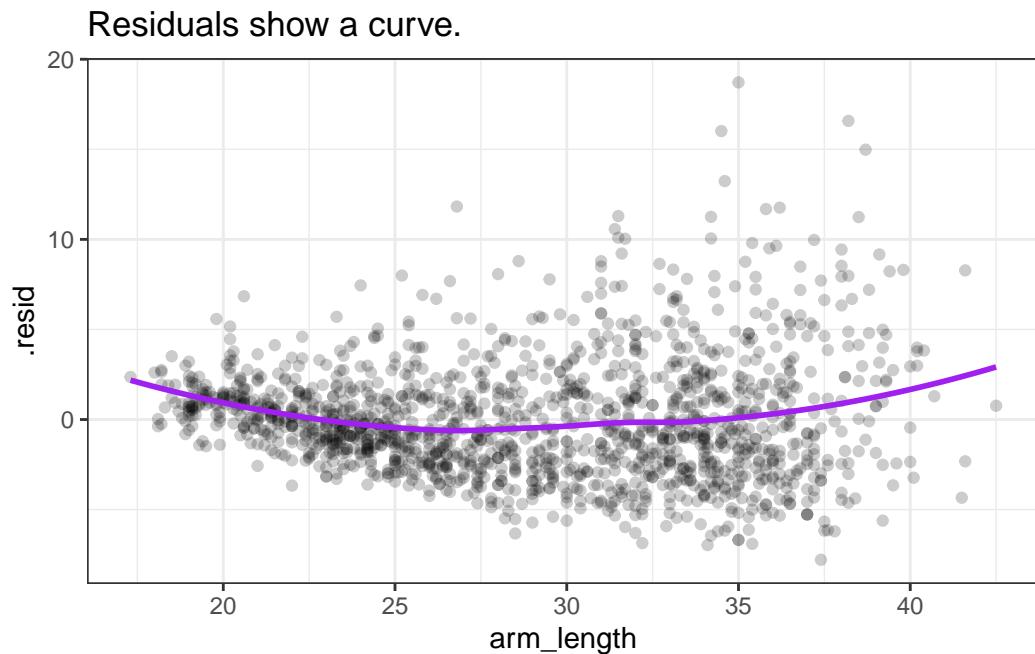
```

OK. So the residuals are now stored in the `.resid` variable. We can create a residual plot, as follows.

```

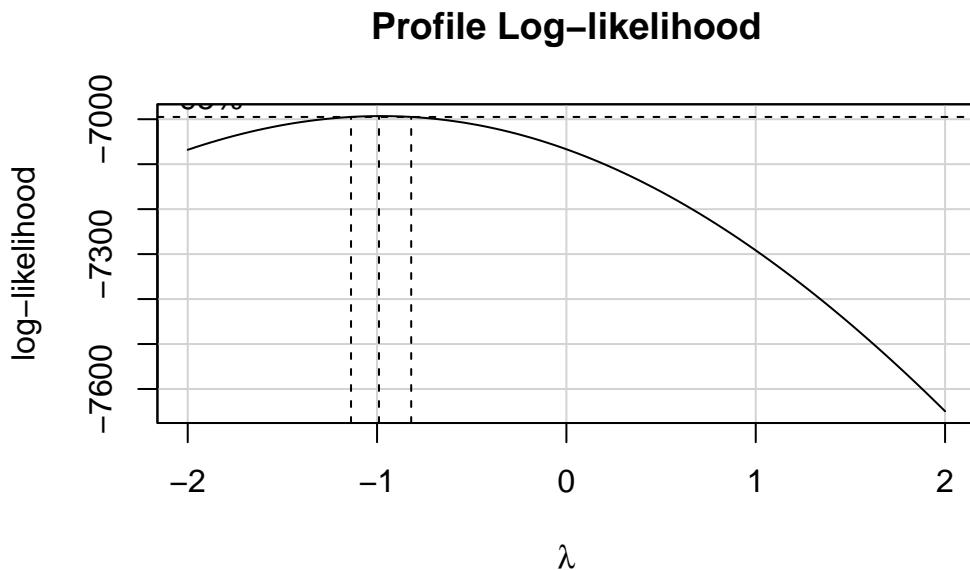
ggplot(nnyfs_c_aug1, aes(x = arm_length, y = .resid)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "loess", col = "purple",
              formula = y ~ x, se = FALSE) +
  labs(title = "Residuals show a curve.")

```



12.6.3 Using the Box-Cox approach to identify a transformation

```
library(car)
boxCox(nnyfs_c$arm_circ ~ nnyfs_c$arm_length)
```



```
powerTransform(nnyfs_c$arm_circ ~ nnyfs_c$arm_length)
```

```
Estimated transformation parameter
Y1
-0.9783135
```

This suggests that we should transform the `arm_circ` data by taking its inverse (power = -1.) Let's take a look at that result.

12.6.4 Plots after Inverse Transformation

Let's build (on the left) the revised scatterplot and (on the right) the revised residual plot after transforming the outcome (`arm_circ`) by taking its inverse.

```

nnyfs_c <- nnyfs_c |>
  mutate(inv_arm_circ = 1/arm_circ)

p1 <- ggplot(nnyfs_c, aes(x = arm_length, y = inv_arm_circ)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "loess", formula = y ~ x,
              se = FALSE, color = "blue") +
  geom_smooth(method = "lm", formula = y ~ x,
              se = FALSE, color = "red") +
  labs(title = "Transformation reduces curve")

m2 <- lm(inv_arm_circ ~ arm_length, data = nnyfs_c)

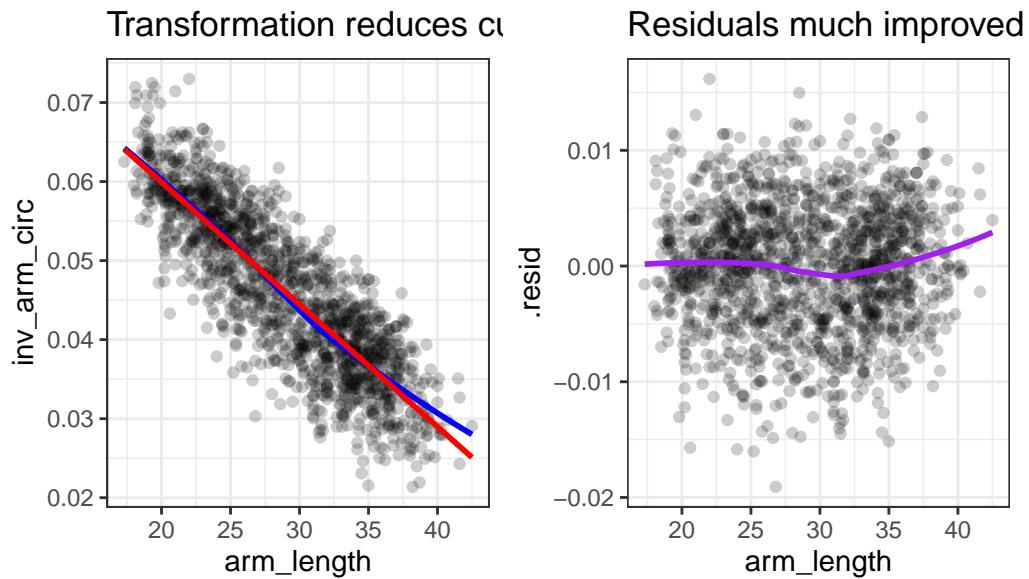
nnyfs_c_aug2 <- augment(m2, data = nnyfs_c)

p2 <- ggplot(nnyfs_c_aug2, aes(x = arm_length, y = .resid)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "loess", col = "purple",
              formula = y ~ x, se = FALSE) +
  labs(title = "Residuals much improved")

p1 + p2 +
  plot_annotation(title = "Evaluating the Inverse Transformation")

```

Evaluating the Inverse Transformation



13 Studying Crab Claws

For our next example, we'll consider a study from zoology, specifically carcinology - the study of crustaceans. My source for these data is Chapter 7 in Ramsey and Schafer (2002) which drew the data from a figure in Yamada and Boulding (1998).

13.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(janitor)
library(broom)
library(knitr)
library(tidyverse)

theme_set(theme_bw())
```

We will also use the `describe` function from the `psych` package.

13.2 The Data

The available data are the mean closing forces (in Newtons) and the propodus heights (mm) of the claws on 38 crabs that came from three different species. The *propodus* is the segment of the crab's clawed leg with an immovable finger and palm.

This was part of a study of the effects that predatory intertidal crab species have on populations of snails. The three crab species under study are:

- 14 *Hemigrapsus nudus*, also called the **purple shore crab** (14 crabs)
- 12 *Lophopanopeus bellus*, also called the **black-clawed pebble crab**, and
- 12 *Cancer productus*, one of several species of **red rock crabs** (12)

```
crabs <- read_csv("data/crabs.csv")
```

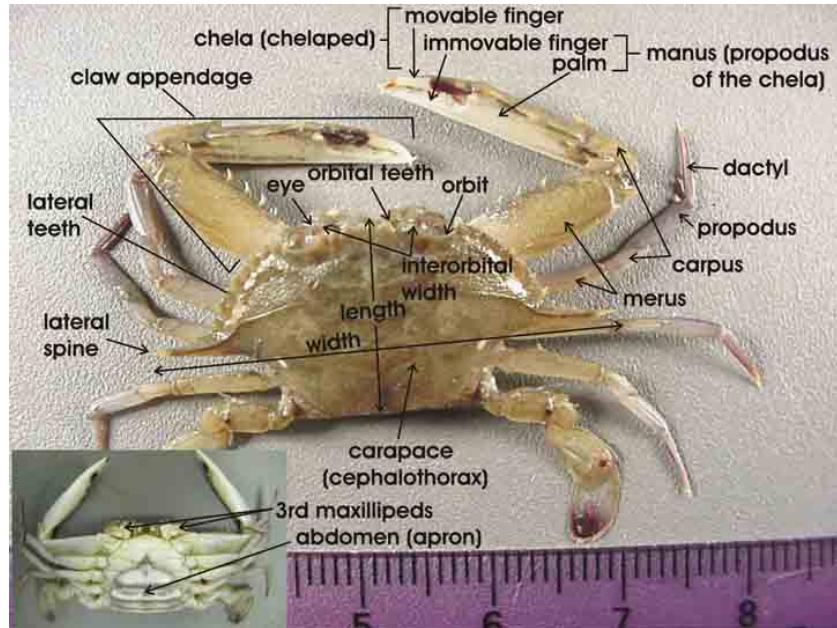


Figure 13.1: Source: <http://txmarspecies.tamug.edu/crustglossary.cfm>

Rows: 38 Columns: 4

-- Column specification -----

Delimiter: ","

chr (1): species

dbl (3): crab, force, height

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

crabs

#	A tibble: 38 x 4	crab	species	force	height
		<dbl>	<chr>	<dbl>	<dbl>
1	1 Hemigrapsus nudus	4	8		
2	2 Lophopanopeus bellus	15.1	7.9		
3	3 Cancer productus	5	6.7		
4	4 Lophopanopeus bellus	2.9	6.6		
5	5 Hemigrapsus nudus	3.2	5		
6	6 Hemigrapsus nudus	9.5	7.9		
7	7 Cancer productus	22.5	9.4		

```

8      8 Hemigrapsus nudus      7.4     8.3
9      9 Cancer productus     14.6    11.2
10     10 Lophopanopeus bellus   8.7     8.6
# ... with 28 more rows
# i Use `print(n = ...)` to see more rows

```

The `species` information is stored here as a character variable. How many different crabs are we talking about in each `species`?

```
crabs |> tabyl(species)
```

	species	n	percent
Cancer productus	12	0.3157895	
Hemigrapsus nudus	14	0.3684211	
Lophopanopeus bellus	12	0.3157895	

As it turns out, we're going to want to treat the `species` information as a **factor** with three levels, rather than as a character variable.

```
crabs <- crabs |>
  mutate(species = factor(species))
```

Here's a quick summary of the data. Take care to note the useless results for the first two variables. At least the function flags with a * those variables it thinks are non-numeric.

```
psych::describe(crabs)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
crab	1	38	19.50	11.11	19.50	19.50	14.08	1	38.0	37.0	0.00	-1.30
species*	2	38	2.00	0.81	2.00	2.00	1.48	1	3.0	2.0	0.00	-1.50
force	3	38	12.13	8.98	8.70	11.53	9.04	2	29.4	27.4	0.47	-1.25
height	4	38	8.81	2.23	8.25	8.78	2.52	5	13.1	8.1	0.19	-1.14
			se									
crab			1.80									
species*			0.13									
force			1.46									
height			0.36									

Actually, we're more interested in these results after grouping by species.

```

crabs |>
  group_by(species) |>
  summarise(n = n(), median(force), median(height))

# A tibble: 3 x 4
  species           n `median(force)` `median(height)`
  <fct>         <int>        <dbl>        <dbl>
1 Cancer productus     12        19.7       11.0
2 Hemigrapsus nudus    14        3.7        7.9
3 Lophopanopeus bellus   12       14.8       8.15

```

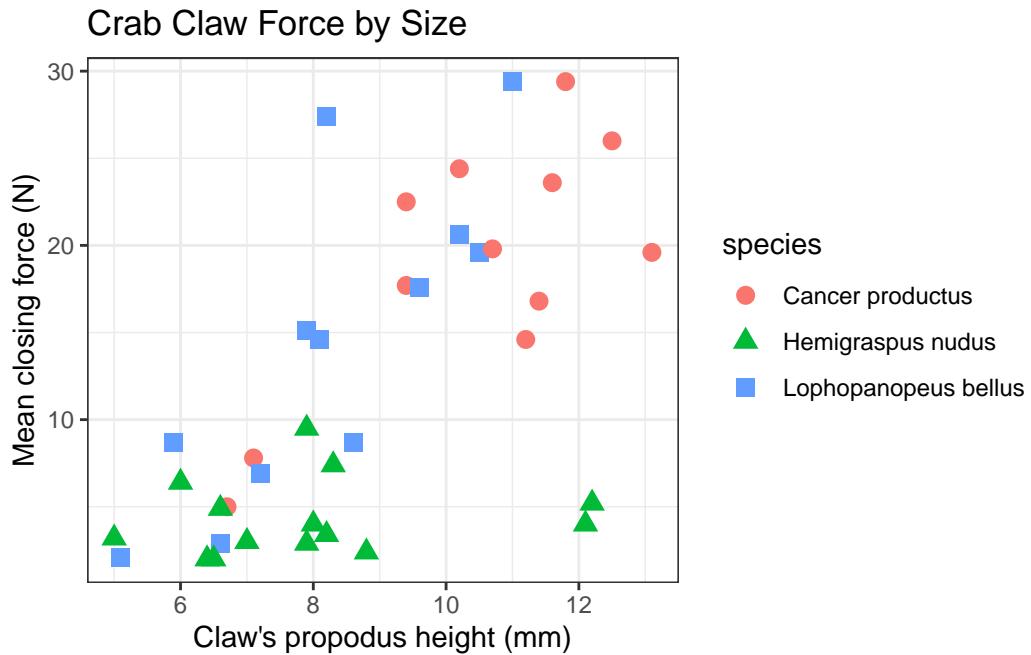
13.3 Association of Size and Force

Suppose we want to describe force on the basis of height, across all 38 crabs. We'll add titles and identify the three species of crab, using shape and color.

```

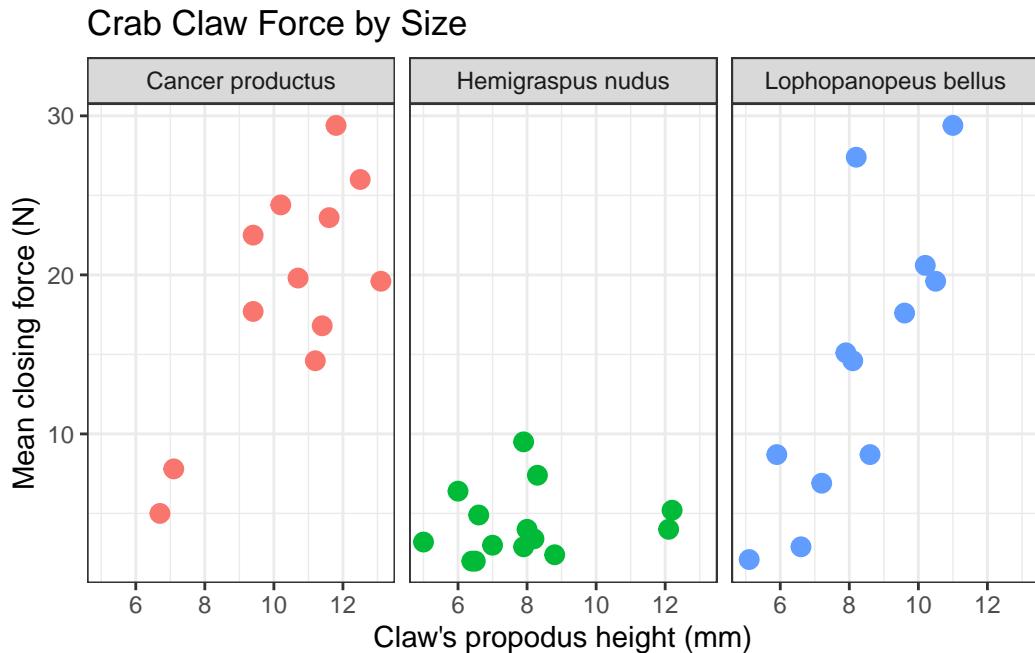
ggplot(crabs, aes(x = height, y = force, color = species, shape = species)) +
  geom_point(size = 3) +
  labs(title = "Crab Claw Force by Size",
       x = "Claw's propodus height (mm)", y = "Mean closing force (N)") +
  theme_bw()

```



A faceted plot for each species really highlights the difference in force between the *Hemigrapsus nudus* and the other two species of crab.

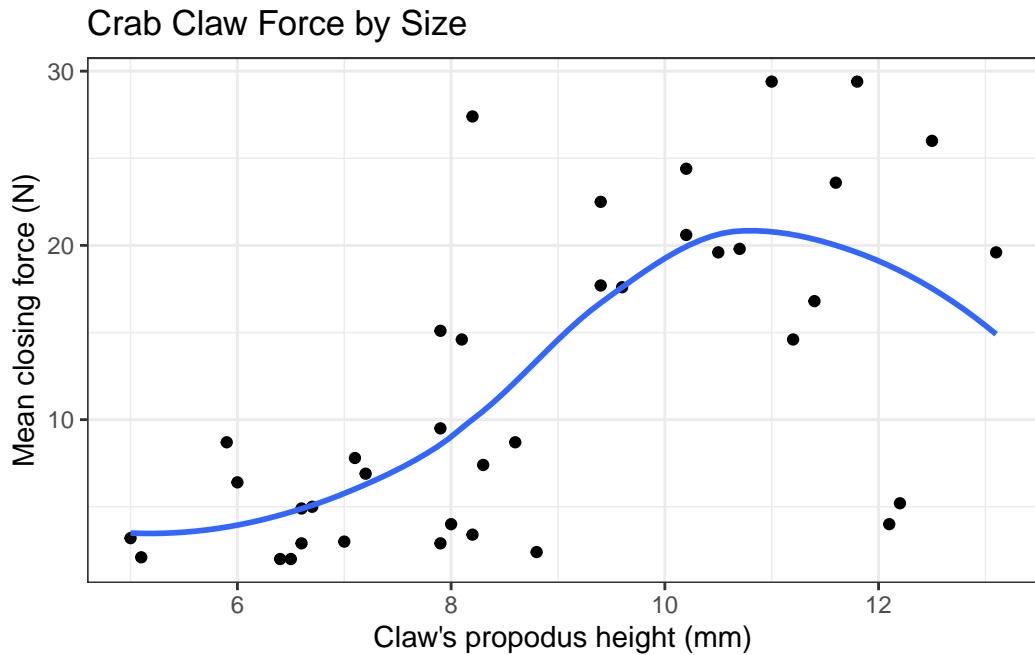
```
ggplot(crabs, aes(x = height, y = force, color = species)) +
  geom_point(size = 3) +
  facet_wrap(~ species) +
  guides(color = "none") +
  labs(title = "Crab Claw Force by Size",
       x = "Claw's propodus height (mm)", y = "Mean closing force (N)") +
  theme_bw()
```



13.4 The loess smooth

We can obtain a smoothed curve (using several different approaches) to summarize the pattern presented by the data in any scatterplot. For instance, we might build such a plot for the complete set of 38 crabs, adding in a non-linear smooth function (called a loess smooth.)

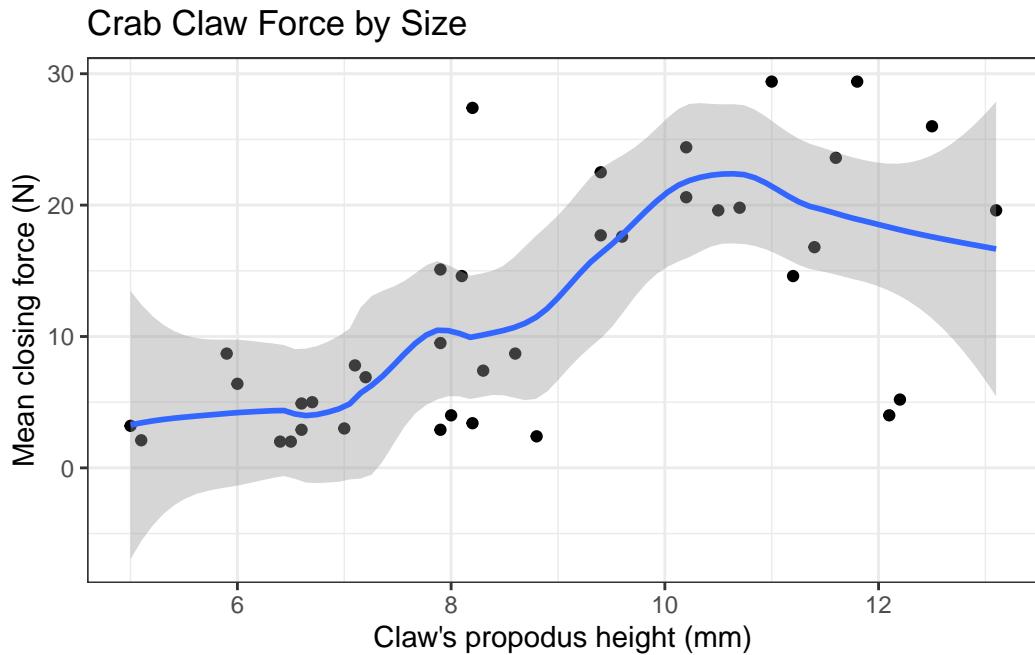
```
ggplot(crabs, aes(x = height, y = force)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE, formula = y ~ x) +
  labs(title = "Crab Claw Force by Size",
       x = "Claw's propodus height (mm)", y = "Mean closing force (N)")
```



As we have discussed previously, a **loess smooth** fits a curve to data by tracking (at point x) the points within a neighborhood of point x , with more emphasis given to points near x . It can be adjusted by tweaking the `span` and `degree` parameters.

In addition to the curve, smoothing procedures can also provide confidence intervals around their main fitted line. Consider the following plot of the `crabs` information, which adjusts the `span` (from its default of 0.75) and also adds in the confidence intervals.

```
ggplot(crabs, aes(x = height, y = force)) +
  geom_point() +
  geom_smooth(method = "loess", formula = y ~ x, span = 0.5, se = TRUE) +
  labs(title = "Crab Claw Force by Size",
       x = "Claw's propodus height (mm)", y = "Mean closing force (N)")
```

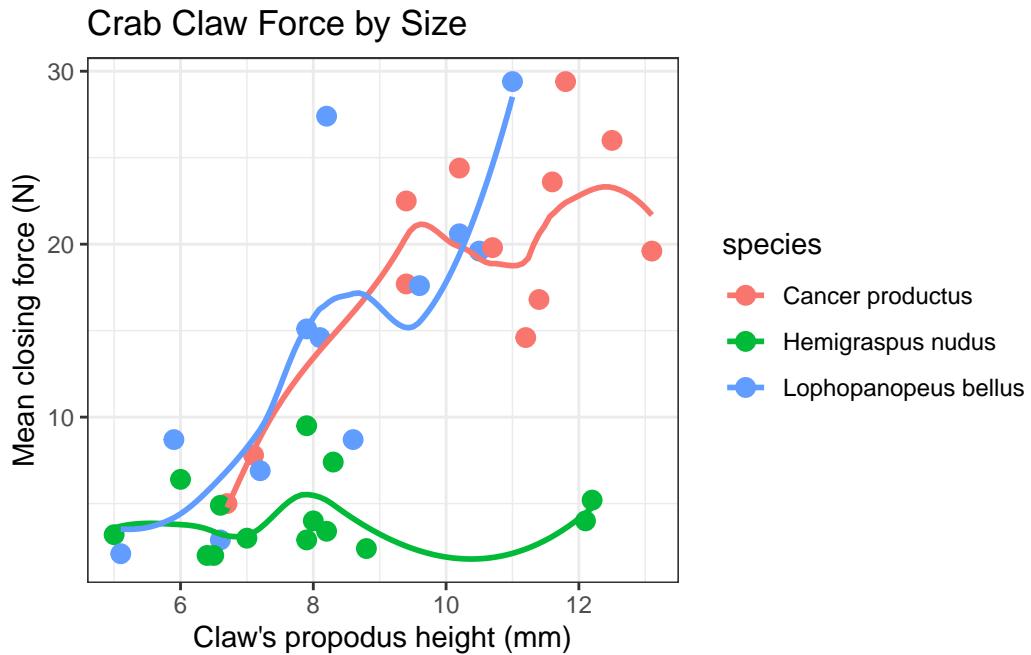


By reducing the size of the span, our resulting picture shows a much less smooth function than we generated previously.

13.4.1 Smoothing within Species

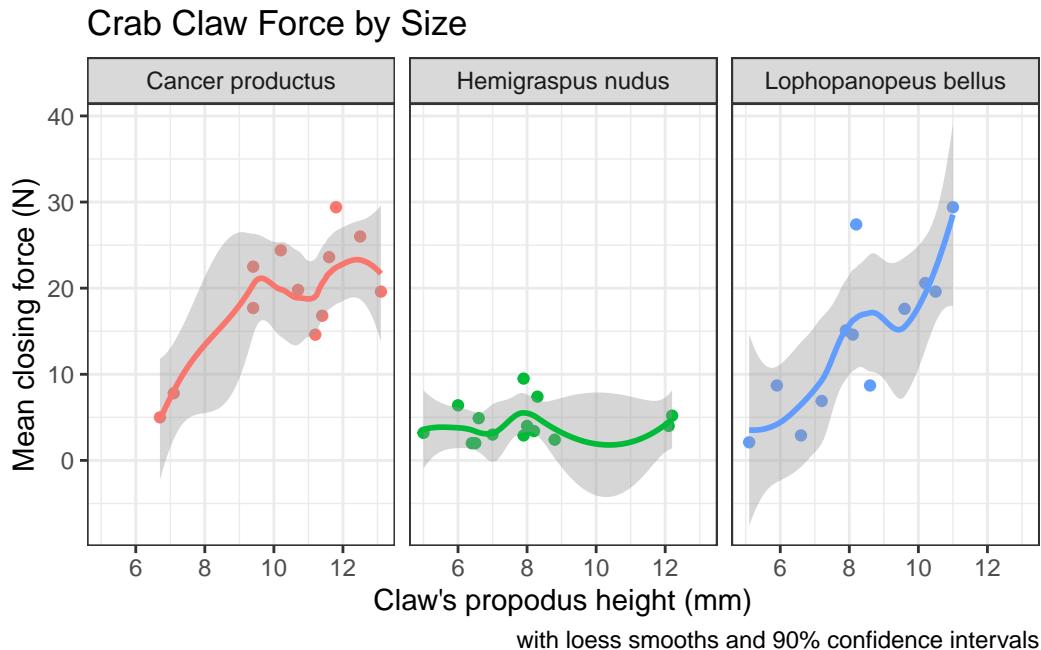
We can, of course, produce the plot above with separate smooths for each of the three species of crab.

```
ggplot(crabs, aes(x = height, y = force, group = species, color = species)) +
  geom_point(size = 3) +
  geom_smooth(method = "loess", formula = y ~ x, se = FALSE) +
  labs(title = "Crab Claw Force by Size",
       x = "Claw's propodus height (mm)", y = "Mean closing force (N)")
```



If we want to add in the confidence intervals (here I'll show them at 90% rather than the default of 95%) then this plot should be faceted. Note that by default, what is displayed when `se = TRUE` are 95% prediction intervals - the `level` function in `stat_smooth` [which can be used in place of `geom_smooth`] is used here to change the coverage percentage from 95% to 90%.

```
ggplot(crabs, aes(x = height, y = force, group = species, color = species)) +
  geom_point() +
  stat_smooth(method = "loess", formula = y ~ x, level = 0.90, se = TRUE) +
  guides(color = "none") +
  labs(title = "Crab Claw Force by Size",
       caption = "with loess smooths and 90% confidence intervals",
       x = "Claw's propodus height (mm)", y = "Mean closing force (N)") +
  facet_wrap(~ species)
```



More on these and other confidence intervals later, especially in part B.

13.5 Fitting a Linear Regression Model

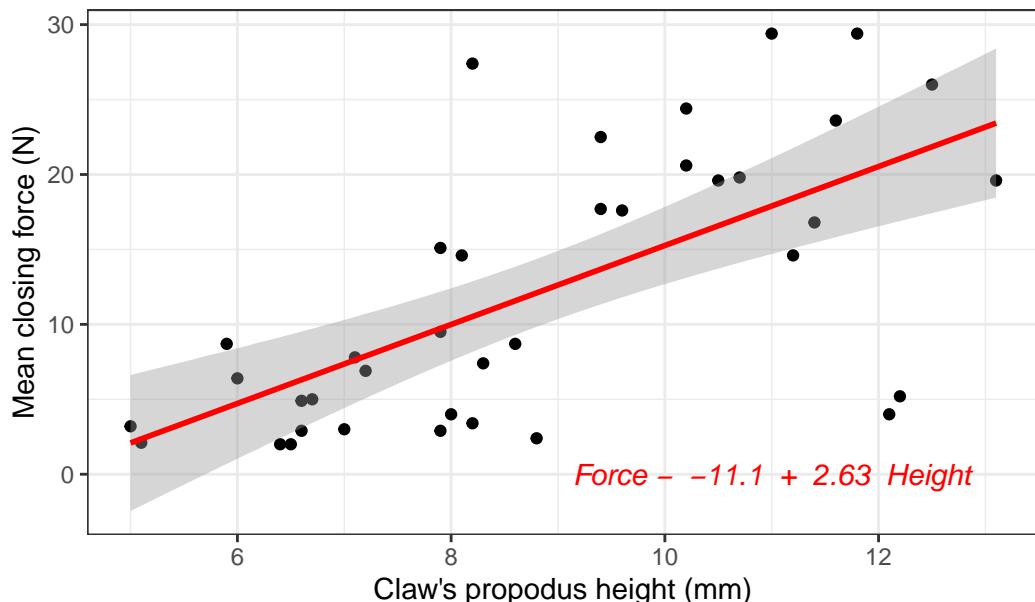
Suppose we plan to use a simple (least squares) linear regression model to describe force as a function of height. Is a least squares model likely to be an effective choice here?

The plot below shows the regression line predicting closing force as a function of propodus height. Here we annotate the plot to show the actual fitted regression line, which required fitting it with the `lm` statement prior to developing the graph.

```
mod <- lm(force ~ height, data = crabs)

ggplot(crabs, aes(x = height, y = force)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, color = "red") +
  labs(title = "Crab Claw Force by Size with Linear Regression Model",
       x = "Claw's propodus height (mm)", y = "Mean closing force (N)") +
  annotate("text", x = 11, y = 0, color = "red", fontface = "italic",
           label = paste( "Force = ", signif(coef(mod)[1],3), " + ",
                         signif(coef(mod)[2],3), " Height" ))
```

Crab Claw Force by Size with Linear Regression Model



```
rm(mod)
```

The **lm** function, again, specifies the linear model we fit to predict force using height. Here's the summary.

```
summary(lm(force ~ height, data = crabs))
```

```
Call:  
lm(formula = force ~ height, data = crabs)
```

Residuals:

Min	1Q	Median	3Q	Max
-16.7945	-3.8113	-0.2394	4.1444	16.8814

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-11.0869	4.6224	-2.399	0.0218 *
height	2.6348	0.5089	5.177	8.73e-06 ***

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

```
Residual standard error: 6.892 on 36 degrees of freedom
Multiple R-squared:  0.4268,    Adjusted R-squared:  0.4109
F-statistic: 26.8 on 1 and 36 DF,  p-value: 8.73e-06
```

Again, the key things to realize are:

- The outcome variable in this model is **force**, and the predictor variable is **height**.
- The straight line model for these data fitted by least squares is $\text{force} = -11.1 + 2.63 \text{ height}$.
- The slope of height is positive, which indicates that as height increases, we expect that force will also increase. Specifically, we expect that for every additional mm of height, the force will increase by 2.63 Newtons.
- The multiple R-squared (squared correlation coefficient) is 0.427, which implies that 42.7% of the variation in force is explained using this linear model with height. It also implies that the Pearson correlation between force and height is the square root of 0.427, or 0.653.

13.6 Is a Linear Model Appropriate?

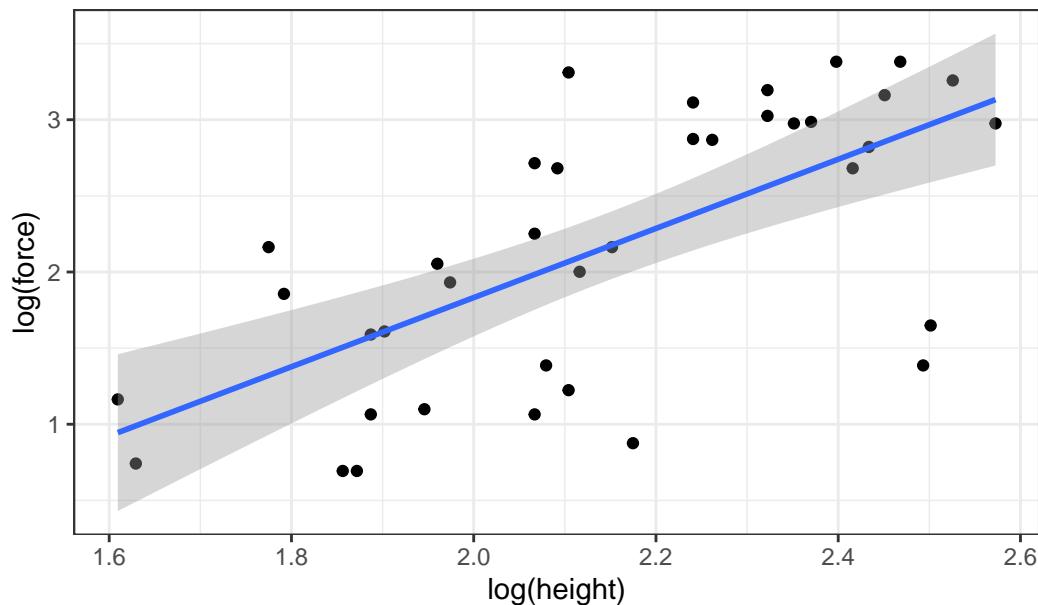
The zoology (at least as described in Ramsey and Schafer (2002)) suggests that the actual nature of the relationship would be represented by a log-log relationship, where the log of force is predicted by the log of height.

This log-log model is an appropriate model when we think that percentage increases in X (height, here) lead to constant percentage increases in Y (here, force).

To see the log-log model in action, we plot the log of force against the log of height. We could use either base 10 (`log10` in R) or natural (`log` in R) logarithms.

```
ggplot(crabs, aes(x = log(height), y = log(force))) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Log-Log Model for Crabs data")
```

Log–Log Model for Crabs data



The correlations between the raw force and height and between their logarithms turn out to be quite similar, and because the log transformation is monotone in these data, there's actually no change at all in the Spearman correlations.

Correlation of	Pearson r	Spearman r
force and height	0.653	0.657
log(force) and log(height)	0.662	0.657

13.6.1 The log-log model

```
crab_loglog <- lm(log(force) ~ log(height), data = crabs)
summary(crab_loglog)
```

Call:
`lm(formula = log(force) ~ log(height), data = crabs)`

Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

```

-1.5657 -0.4450  0.1884  0.4798  1.2422

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.7104     0.9251  -2.930  0.00585 **
log(height)  2.2711     0.4284   5.302 5.96e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6748 on 36 degrees of freedom
Multiple R-squared:  0.4384,    Adjusted R-squared:  0.4228
F-statistic: 28.11 on 1 and 36 DF,  p-value: 5.96e-06

```

Our regression equation is $\log(\text{force}) = -2.71 + 2.27 \log(\text{height})$.

So, for example, if we found a crab with propodus height = 10 mm, our prediction for that crab's claw force (in Newtons) based on this log-log model would be...

- $\log(\text{force}) = -2.71 + 2.27 \log(10)$
- $\log(\text{force}) = -2.71 + 2.27 \times 2.3025851$
- $\log(\text{force}) = 2.5190953$
- and so predicted force = $\exp(2.5190953) = 12.4173582$ Newtons, which, naturally, we would round to 12.4 Newtons to match the data set's level of precision.

13.6.2 How does this compare to our original linear model?

```

crab_linear <- lm(force ~ height, data = crabs)

summary(crab_linear)

```

Call:
`lm(formula = force ~ height, data = crabs)`

Residuals:

Min	1Q	Median	3Q	Max
-16.7945	-3.8113	-0.2394	4.1444	16.8814

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-11.0869	4.6224	-2.399	0.0218 *

```

height          2.6348      0.5089    5.177 8.73e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.892 on 36 degrees of freedom
Multiple R-squared:  0.4268,   Adjusted R-squared:  0.4109
F-statistic:  26.8 on 1 and 36 DF,  p-value: 8.73e-06

```

The linear regression equation is force = $-11.1 + 2.63 \text{ height}$.

So, for example, if we found a crab with propodus height = 10 mm, our prediction for that crab's claw force (in Newtons) based on this linear model would be...

- force = $-11.0869025 + 2.6348232 \times 10$
- force = $-11.0869025 + 26.3482321$
- so predicted force = 15.2613297, which we would round to 15.3 Newtons.

So, it looks like the two models give meaningfully different predictions.

13.7 Making Predictions with a Model

The `broom` package's `augment` function provides us with a consistent method for obtaining predictions (also called fitted values) for a new crab or for our original data. Suppose we want to predict the `force` level for two new crabs: one with height = 10 mm, and another with height = 12 mm.

```

newcrab <- tibble(crab = c("Crab_A", "Crab_B"), height = c(10, 12))

augment(crab_linear, newdata = newcrab)

# A tibble: 2 x 3
  crab    height .fitted
  <chr>    <dbl>    <dbl>
1 Crab_A     10     15.3
2 Crab_B     12     20.5

```

Should we want to obtain a prediction interval, we can use the `predict` function:

```
predict(crab_linear, newdata = newcrab, interval = "prediction", level = 0.95)
```

```
    fit      lwr      upr
1 15.26133 1.048691 29.47397
2 20.53098 5.994208 35.06774
```

We'd interpret this result as saying that the linear model's predicted force associated with a single new crab claw with propodus height 10 mm is 15.3 Newtons, and that a 95% prediction interval for the true value of such a force for such a claw is between 1.0 and 29.5 Newtons. More on prediction intervals later.

13.7.1 Predictions After a Transformation

We can also get predictions from the log-log model. The default choice is a 95% prediction interval.

```
predict(crab_loglog, newdata = newcrab, interval = "prediction")
```

```
    fit      lwr      upr
1 2.519095 1.125900 3.912291
2 2.933174 1.515548 4.350800
```

Of course, these predictions describe the `log(force)` for such a crab claw. To get the prediction in terms of simple force, we'd need to back out of the logarithm, by exponentiating our point estimate and the prediction interval endpoints.

```
exp(predict(crab_loglog, newdata = newcrab, interval = "prediction"))
```

```
    fit      lwr      upr
1 12.41736 3.082989 50.01341
2 18.78716 4.551916 77.54044
```

We'd interpret this result as saying, for the first new crab, that the log-log model's predicted force associated with a single new crab claw with propodus height 10 mm is 12.4 Newtons, and that a 95% prediction interval for the true value of such a force for such a claw is between 3.1 and 50.0 Newtons.

13.7.2 Comparing Model Predictions

Suppose we wish to build a plot of force vs height with a straight line for the linear model's predictions, and a new curve for the log-log model's predictions, so that we can compare and contrast the implications of the two models on a common scale. The `predict` function, when not given a new data frame, will use the existing predictor values that are in our `crabs` data. Such predictions are often called fitted values.

To put the two sets of predictions on the same scale despite the differing outcomes in the two models, we'll exponentiate the results of the log-log model, and build a little data frame containing the heights and the predicted forces from that model.

```
loglogdat <- tibble(height = crabs$height, force = exp(predict(crab_loglog)))
```

A cleaner way to do this might be to use the `augment` function directly from `broom`:

```
augment(crab_loglog)
```

```
# A tibble: 38 x 7
`log(force)` `log(height)` .fitted   .hat   .sigma   .cooksdi .std.resid
<dbl>        <dbl>     <dbl>   <dbl>    <dbl>     <dbl>      <dbl>
1       1.39      2.08    2.01  0.0280  0.676 1.28e- 2  -0.941
2       2.71      2.07    1.98  0.0287  0.673 1.79e- 2   1.10 
3       1.61      1.90    1.61  0.0499  0.684 8.06e-10 -0.000175
4       1.06      1.89    1.58  0.0530  0.679 1.69e- 2  -0.778
5       1.16      1.61    0.945 0.142   0.683 1.01e- 2   0.349 
6       2.25      2.07    1.98  0.0287  0.683 2.39e- 3   0.402 
7       3.11      2.24    2.38  0.0301  0.673 1.90e- 2   1.11 
8       2.00      2.12    2.10  0.0266  0.684 2.75e- 4  -0.142 
9       2.68      2.42    2.78  0.0561  0.684 6.30e- 4  -0.146 
10      2.16      2.15    2.18  0.0263  0.684 5.34e- 6  -0.0199
# ... with 28 more rows
# i Use `print(n = ...)` to see more rows
```

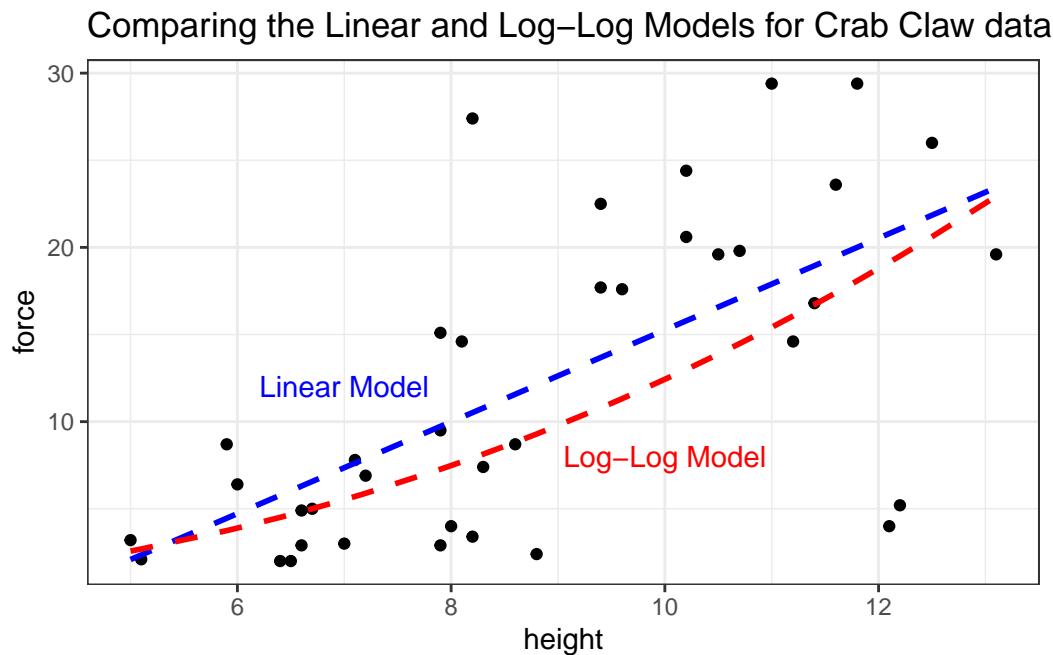
Now, we're ready to use the `geom_smooth` approach to plot the linear fit, and `geom_line` (which also fits curves) to display the log-log fit.

```
ggplot(crabs, aes(x = height, y = force)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE,
              formula = y ~ x, col="blue", linetype = 2) +
  geom_line(data = loglogdat, col = "red", linetype = 2, size = 1) +
```

```

annotate("text", 7, 12, label = "Linear Model", col = "blue") +
annotate("text", 10, 8, label = "Log-Log Model", col = "red") +
labs(title = "Comparing the Linear and Log-Log Models for Crab Claw data")

```



Based on these 38 crabs, we see some modest differences between the predictions of the two models, with the log-log model predicting generally lower closing force for a given propodus height than would be predicted by a linear model.

14 Dehydration Recovery

14.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(knitr)
library(broom)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

We will also use the `ggpairs` function from the `GGally` package, and the `favstats` function from the `mosaic` package.

14.2 The Data

The `hydrate` data describe the degree of recovery that takes place 90 minutes following treatment of moderate to severe dehydration, for 36 children diagnosed at a hospital's main pediatric clinic.

Upon diagnosis and study entry, patients were treated with an electrolytic solution at one of seven `dose` levels (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0 mEq/l) in a frozen, flavored, ice popsicle. The degree of rehydration was determined using a subjective scale based on physical examination and parental input, converted to a 0 to 100 point scale, representing the percent of recovery (`recov.score`). Each child's `age` (in years) and `weight` (in pounds) are also available.

First, we'll check ranges (and for missing data) in the `hydrate` file.

```
hydrate <- read_csv("data/hydrate.csv")

summary(hydrate)
```

```

      id      recov.score      dose       age
Min.   : 1.00  Min.   :44.00  Min.   :0.000  Min.   : 3.000
1st Qu.: 9.75  1st Qu.:61.50  1st Qu.:1.000  1st Qu.: 5.000
Median :18.50  Median :71.50  Median :1.500  Median : 6.500
Mean   :18.50  Mean   :71.56  Mean   :1.569  Mean   : 6.667
3rd Qu.:27.25  3rd Qu.:80.00  3rd Qu.:2.500  3rd Qu.: 8.000
Max.   :36.00  Max.   :100.00  Max.   :3.000  Max.   :11.000

      weight
Min.   :22.00
1st Qu.:34.50
Median :47.50
Mean   :46.89
3rd Qu.:57.25
Max.   :76.00

```

There are no missing values, and all of the ranges make sense. There are no especially egregious problems to report.

14.3 A Scatterplot Matrix

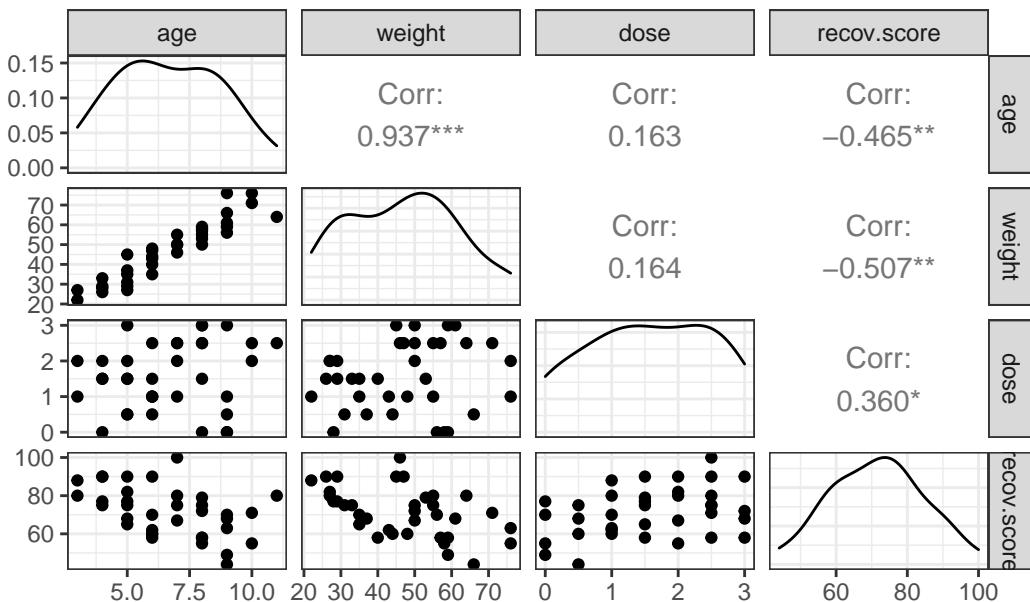
Next, we'll use a scatterplot matrix to summarize relationships between the outcome `recov.score` and the key predictor `dose` as well as the ancillary predictors `age` and `weight`, which are of less interest, but are expected to be related to our outcome. The one below uses the `ggpairs` function in the `GGally` package, as introduced in Part A of the Notes. We place the outcome in the bottom row, and the key predictor immediately above it, with `age` and `weight` in the top rows, using the `select` function within the 'ggpairs' call.

```
GGally::ggpairs(dplyr::select(hydrate, age, weight, dose, recov.score),
                 title = "Scatterplot Matrix for hydrate data")
```

Registered S3 method overwritten by 'GGally':

```
method from
+.gg   ggplot2
```

Scatterplot Matrix for hydrate data



What can we conclude here?

- It looks like `recov.score` has a moderately strong negative relationship with both `age` and `weight` (with correlations in each case around -0.5), but a positive relationship with `dose` (correlation = 0.36).
- The distribution of `recov.score` looks to be pretty close to Normal. No potential predictors (`age`, `weight` and `dose`) show substantial non-Normality.
- `age` and `weight`, as we'd expect, show a very strong and positive linear relationship, with $r = 0.94$
- Neither `age` nor `weight` shows a meaningful relationship with `dose`. ($r = 0.16$)

14.4 Are the recovery scores well described by a Normal model?

Next, we'll do a more thorough graphical summary of our outcome, recovery score.

```
p1 <- ggplot(hydrate, aes(sample = recov.score)) +
  geom_qq(col = '#440154') + geom_qq_line(col = "red") +
  theme(aspect.ratio = 1) +
  labs(title = "Normal Q-Q plot: hydrate")

p2 <- ggplot(hydrate, aes(x = recov.score)) +
```

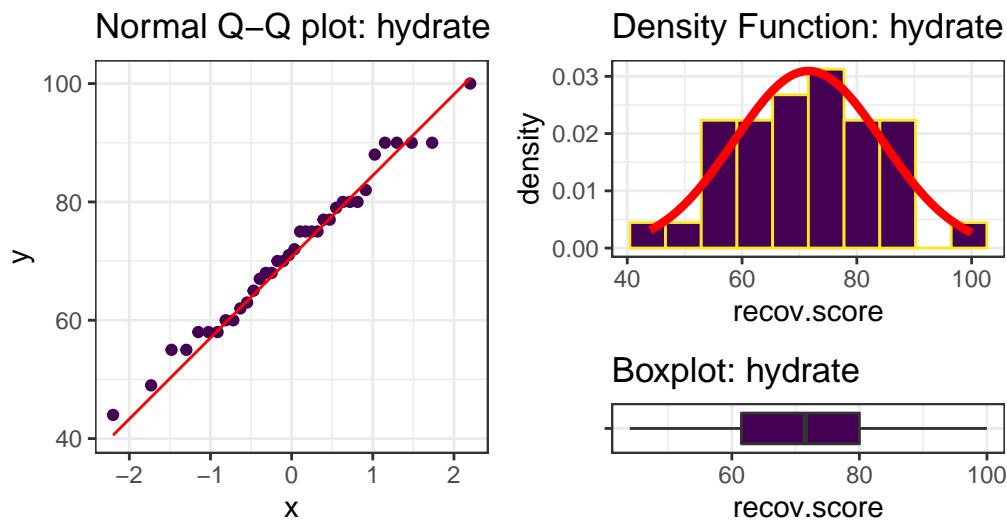
```

geom_histogram(aes(y = stat(density)),
               bins = 10, fill = '#440154', col = '#FDE725') +
stat_function(fun = dnorm,
              args = list(mean = mean(hydrate$recov.score),
                          sd = sd(hydrate$recov.score)),
              col = "red", lwd = 1.5) +
labs(title = "Density Function: hydrate")

p3 <- ggplot(hydrate, aes(x = recov.score, y = "")) +
geom_boxplot(fill = '#440154', outlier.color = '#440154') +
labs(title = "Boxplot: hydrate", y = "")

p1 + (p2 / p3 + plot_layout(heights = c(4,1)))

```



```
mosaic::favstats(~ recov.score, data = hydrate) |> kable(digits = 1)
```

```

Registered S3 method overwritten by 'mosaic':
method                      from
fortify.SpatialPolygonsDataFrame ggplot2

```

min	Q1	median	Q3	max	mean	sd	n	missing
44	61.5	71.5	80	100	71.6	12.9	36	0

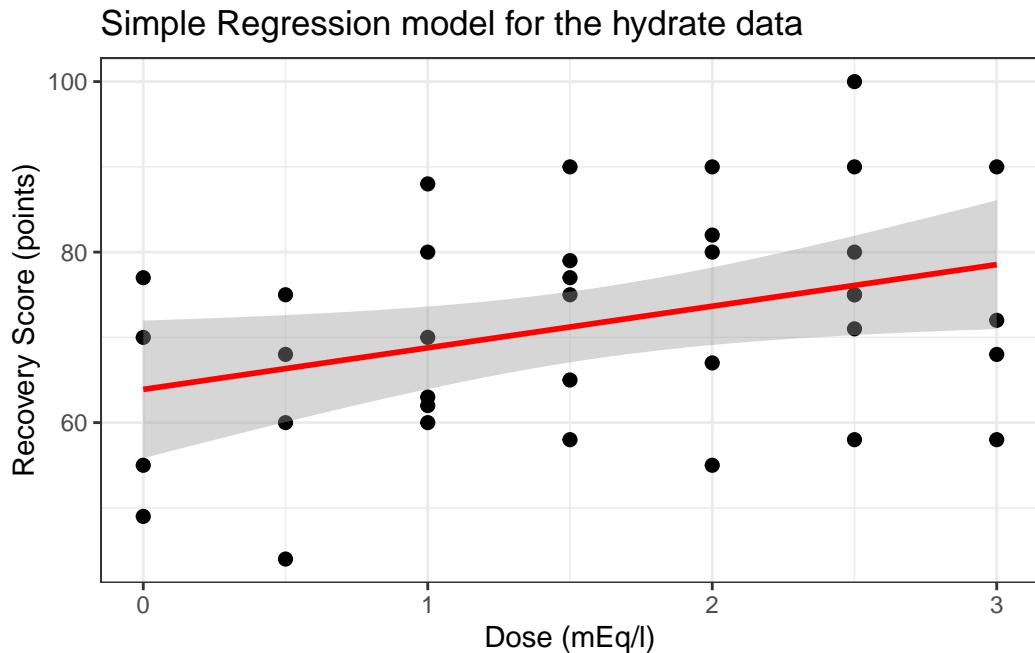
I see no serious problems with assuming Normality for these recovery scores. Our outcome variable doesn't in any way *need* to follow a Normal distribution, but it's nice when it does, because summaries involving means and standard deviations make sense.

14.5 Simple Regression: Using Dose to predict Recovery

To start, consider a simple (one predictor) regression model using `dose` alone to predict the % Recovery (`recov.score`). Ignoring the `age` and `weight` covariates, what can we conclude about this relationship?

14.6 The Scatterplot, with fitted Linear Model

```
ggplot(hydrate, aes(x = dose, y = recov.score)) +
  geom_point(size = 2) +
  geom_smooth(method = "lm", formula = y ~ x, col = "red") +
  labs(title = "Simple Regression model for the hydrate data",
       x = "Dose (mEq/l)", y = "Recovery Score (points)")
```



14.7 The Fitted Linear Model

To obtain the fitted linear regression model, we use the `lm` function:

```
m1 <- lm(recov.score ~ dose, data = hydrate)

tidy(m1) |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	63.90	3.97	16.09	0.00
dose	4.88	2.17	2.25	0.03

So, our fitted regression model (prediction model) is `recov.score = 63.9 + 4.88 dose`.

14.7.1 Confidence Intervals

We can obtain confidence intervals around the coefficients of our fitted model with `tidy`, too.

```
tidy(m1, conf.int = TRUE, conf.level = 0.90) |> kable(digits = 2)
```

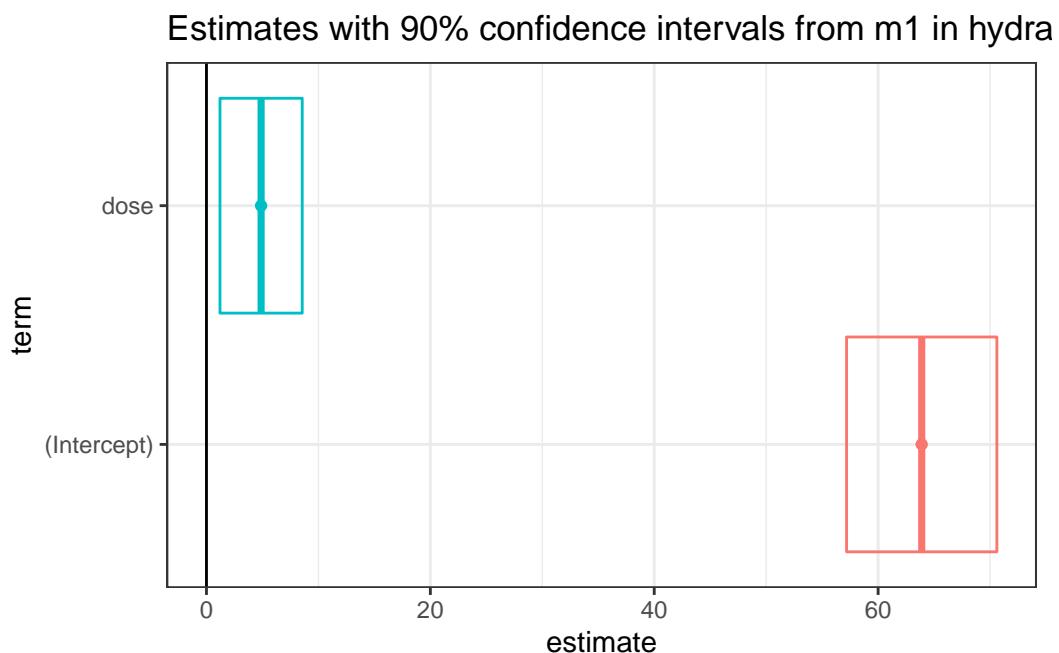
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	63.90	3.97	16.09	0.00	57.18	70.61
dose	4.88	2.17	2.25	0.03	1.21	8.55

So, our 90% confidence interval for the slope of dose ranges from 1.21 to 8.55.

14.8 Coefficient Plots

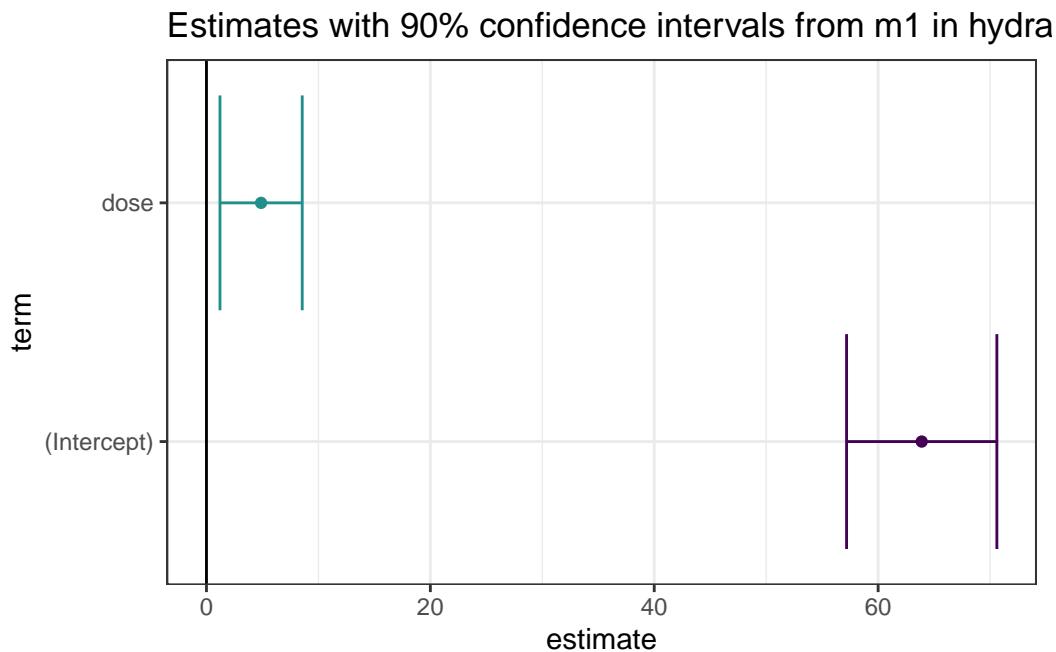
The `tidy` method makes it easy to construct coefficient plots using `ggplot2`.

```
td <- tidy(m1, conf.int = TRUE, conf.level = 0.90)
ggplot(td, aes(x = estimate, y = term, col = term)) +
  geom_point() +
  geom_crossbar(aes(xmin = conf.low, xmax = conf.high)) +
  geom_vline(xintercept = 0) +
  guides(col = "none") +
  labs(title = "Estimates with 90% confidence intervals from m1 in hydrate")
```



Another option would be to use `geom_errorbarh` in this setting, perhaps with a different color scheme...

```
td <- tidy(m1, conf.int = TRUE, conf.level = 0.90)
ggplot(td, aes(x = estimate, y = term, col = term)) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high)) +
  geom_vline(xintercept = 0) +
  scale_color_viridis_d(end = 0.5) +
  guides(col = "none") +
  labs(title = "Estimates with 90% confidence intervals from m1 in hydrate")
```



14.9 The Summary Output

To get a more complete understanding of the fitted model, we'll summarize it.

```
summary(lm(recov.score ~ dose, data = hydrate))
```

Call:

```

lm(formula = recov.score ~ dose, data = hydrate)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.3360 -7.2763  0.0632  8.4233 23.9028 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 63.896     3.970   16.093 <2e-16 ***
dose        4.881     2.172    2.247   0.0313 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.21 on 34 degrees of freedom
Multiple R-squared:  0.1293,    Adjusted R-squared:  0.1037 
F-statistic: 5.047 on 1 and 34 DF,  p-value: 0.03127

```

14.9.1 Model Specification

1. The first part of the output specifies the model that has been fit.
 - Here, we have a simple regression model that predicts `recov.score` on the basis of `dose`.
 - Notice that we're treating `dose` here as a quantitative variable. If we wanted `dose` to be treated as a factor, we'd have specified that in the model.

14.9.2 Residual Summary

2. The second part of the output summarizes the regression **residuals** across the subjects involved in fitting the model.
 - The **residual** is defined as the Actual value of our outcome minus the predicted value of that outcome fitted by the model.
 - In our case, the residual for a given child is their actual `recov.score` minus the predicted `recov.score` according to our model, for that child.
 - The residual summary gives us a sense of how “incorrect” our predictions are for the `hydrate` observations.
 - A positive residual means that the observed value was higher than the predicted value from the linear regression model, so the prediction was too low.
 - A negative residual means that the observed value was lower than the predicted value from the linear regression model, so the prediction was too high.

- The residuals will center near 0 (the ordinary least squares model fitting process is designed so the mean of the residuals will always be zero)
- We hope to see the median of the residuals also be near zero, generally. In this case, the median prediction is 0.06 point too low.
- The minimum and maximum show us the largest prediction errors, made in the subjects used to fit this model.
- Here, we predicted a recovery score that was 22.3 points too high for one patient, and another of our predicted recovery scores was 23.9 points too low.
- The middle half of our predictions were between 8.4 points too low and 7.3 points too high.

14.9.3 Coefficients Output

- The **Coefficients** output begins with a table of the estimated coefficients from the regression equation.
 - Generally, we write a simple regression model as $y = \beta_0 + \beta_1 x$.
 - In the `hydrate` model, we have `recov.score = \beta_0 + \beta_1 dose`.
 - The first column of the table gives the estimated β coefficients for our model
 - Here the estimated intercept $\hat{\beta}_0 = 63.9$
 - The estimated slope of dose $\hat{\beta}_1 = 4.88$
 - Thus, our model is `recov.score = 63.9 + 4.88 dose`

We interpret these coefficients as follows:

- The intercept (63.9) is the predicted `recov.score` for a patient receiving a `dose` of 0 mEq/l of the electrolytic solution.
- The slope (4.88) of the `dose` is the predicted *change* in `recov.score` associated with a 1 mEq/l increase in the dose of electrolytic solution.
 - Essentially, if we have two children like the ones studied here, and we give Roger a popsicle with dose X and Sarah a popsicle with dose X + 1, then this model predicts that Sarah will have a recovery score that is 4.88 points higher than will Roger.
 - From the confidence interval output we saw previously with the function `confint(lm(recov.score ~ dose))`, we are 95% confident that the true slope for `dose` is between (0.47, 9.30) mEq/l. We are also 95% confident that the true intercept is between (55.8, 72.0).

14.9.4 Correlation and Slope

If we like, we can use the `cor` function to specify the Pearson correlation of `recov.score` and `dose`, which turns out to be 0.36. - Note that the `slope` in a simple regression model will

follow the sign of the Pearson correlation coefficient, in this case, both will be positive.

```
hydrate |> select(recov.score, dose) |> cor()
```

```
      recov.score      dose
recov.score    1.000000 0.359528
dose          0.359528 1.000000
```

14.9.5 Coefficient Testing

```
summary(lm(recov.score ~ dose, data = hydrate))
```

Call:

```
lm(formula = recov.score ~ dose, data = hydrate)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.3360	-7.2763	0.0632	8.4233	23.9028

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	63.896	3.970	16.093	<2e-16 ***
dose	4.881	2.172	2.247	0.0313 *

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

Residual standard error: 12.21 on 34 degrees of freedom

Multiple R-squared: 0.1293, Adjusted R-squared: 0.1037

F-statistic: 5.047 on 1 and 34 DF, p-value: 0.03127

Next to each coefficient in the summary regression table is its estimated standard error, followed by the coefficient's t value (the coefficient value divided by the standard error), and the associated two-tailed *p* value for the test of:

- H₀: This coefficient's β value = 0 vs.
- H_A: This coefficient's β value \neq 0.

For the slope coefficient, we can interpret this choice as:

- H₀: This predictor adds no predictive value to the model vs.
- H_A: This predictor adds some predictive value to the model.

In the `hydrate` simple regression model, by running either `tidy` with or just the `confint` function shown below, we can establish a confidence interval for each of the estimated regression coefficients.

```
tidy(m1, conf.int = TRUE, conf.level = 0.95) |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	63.90	3.97	16.09	0.00	55.83	71.96
dose	4.88	2.17	2.25	0.03	0.47	9.30

```
confint(m1, level = .95)
```

	2.5 %	97.5 %
(Intercept)	55.826922	71.964589
dose	0.465695	9.295466

If the slope of dose was in fact zero, then this would mean that knowing the dose information would be of no additional value in predicting the outcome over just guessing the mean of `recov.score` for every subject.

So, since the confidence interval for the slope of dose does not include zero, it appears that there is at least some evidence that the model `m1` is more effective than a model that ignores the `dose` information (and simply predicts the mean of `recov.score` for each subject.) That's not saying much, actually.

14.9.6 Summarizing the Quality of Fit

4. The next part of the regression summary output is a summary of fit quality.

The **residual standard error** estimates the standard deviation of the prediction errors made by the model.

- If assumptions hold, the model will produce residuals that follow a Normal distribution with mean 0 and standard deviation equal to this residual standard error.
 - So we'd expect roughly 95% of our residuals to fall between -2(12.21) and +2(12.21), or roughly -24.4 to +24.4 and that we'd see virtually no residuals outside the range of -3(12.21) to +3(12.21), or roughly -36.6 to +36.6.

- The output at the top of the summary tells us about the observed regression residuals, and that they actually range from -22 to +24.
- In context, it's hard to know whether or not we should be happy about this. On a scale from 0 to 100, rarely missing by more than 24 seems OK to me, but not terrific.
- The **degrees of freedom** here are the same as the denominator degrees of freedom in the ANOVA to follow. The calculation is $n - k$, where n = the number of observations and k is the number of coefficients estimated by the regression (including the intercept and any slopes).
 - Here, there are 36 observations in the model, and we fit $k = 2$ coefficients; the slope and the intercept, as in any simple regression model, so $df = 36 - 2 = 34$.

The multiple R-squared value is usually just referred to as R-squared.

- This is interpreted as the proportion of variation in the outcome variable that has been accounted for by our regression model.
 - Here, we've accounted for just under 13% of the variation in % Recovery using Dose.
- The R in multiple R-squared is the Pearson correlation of `recov.score` and `dose`, which in this case is 0.3595.
 - Squaring this value gives the R-squared for this simple regression.
 - $(0.3595)^2 = 0.129$

R-squared is greedy.

- R-squared will always suggest that we make our models as big as possible, often including variables of dubious predictive value.
- As a result, there are various methods for adjusting or penalizing R-squared so that we wind up with smaller models.
- The **adjusted R-squared** is often a useful way to compare multiple models for the same response.
 - $R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-k}$, where n = the number of observations and k is the number of coefficients estimated by the regression (including the intercept and any slopes).
 - So, in this case, $R_{adj}^2 = 1 - \frac{(1-0.1293)(35)}{34} = 0.1037$
 - The adjusted R-squared value is not, technically, a proportion of anything, but it is comparable across models for the same outcome.
 - The adjusted R-squared will always be less than the (unadjusted) R-squared.

14.9.7 ANOVA F test

5. The last part of the standard summary of a regression model is the overall ANOVA F test.

The hypotheses for this test are:

- H₀: Each of the coefficients in the model (other than the intercept) has $\beta = 0$ vs.
- H_A: At least one regression slope has $\beta \neq 0$

Since we are doing a simple regression with just one predictor, the ANOVA F test hypotheses are exactly the same as the t test for dose:

- H₀: The slope for `dose` has $\beta = 0$ vs.
- H_A: The slope for `dose` has $\beta \neq 0$

In this case, we have an F statistic of 5.05 on 1 and 34 degrees of freedom, yielding $p = 0.03$

This provides some evidence that “something” in our model (here, `dose` is the only predictor) predicts the outcome to a degree beyond that easily attributed to chance alone. This is not actually surprising, nor is it especially interesting. The confidence interval for the slope is definitely more interesting than this.

- In *simple regression* (regression with only one predictor), the t test for the slope (`dose`) always provides the same p value as the ANOVA F test.
 - The F test statistic in a *simple regression* is always by definition just the square of the slope’s t test statistic.
 - Here, $F = 5.047$, and this is the square of $t = 2.247$ from the Coefficients output

This test is basically just a combination of the R-squared value (13%) and the sample size. We don’t learn much from it that’s practically interesting or useful.

14.10 Viewing the complete ANOVA table

We can obtain the complete ANOVA table associated with this particular model, and the details behind this F test using the `anova` function:

```
anova(lm(recov.score ~ dose, data = hydrate))
```

Analysis of Variance Table

```
Response: recov.score
          Df Sum Sq Mean Sq F value Pr(>F)
dose      1 752.2  752.15  5.0473 0.03127 *
Residuals 34 5066.7   149.02
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The R-squared for our regression model is equal to the η^2 for this ANOVA model.
 - If we divide $SS(dose) = 752.2$ by the total sum of squares ($752.2 + 5066.7$), we'll get the multiple R-squared [0.1293]
- Note that this is *not* the same ANOVA model we would get if we treated `dose` as a factor with seven levels, rather than as a quantitative variable.

14.11 Using `glance` to summarize the model's fit

When applied to a linear model, the `glance` function from the `broom` package summarizes 12 characteristics of the model's fit.

Let's look at the eight of these that we've already addressed.

```
glance(m1) |> select(r.squared:df, df.residual, nobs) |>
  kable(digits = c(3, 3, 1, 2, 3, 0, 0, 0))
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual	nobs
0.129	0.104	12.2	5.05	0.031	1	34	36

- We've discussed the R-square value, shown in `r.squared`.
- We've also discussed the adjusted R-square value, in `adj.r.squared`
- `sigma` is the residual standard error.
- `statistic` is the ANOVA F statistic.
- `p.value` is the p value associated with the ANOVA F statistic.
- `df` is the numerator degrees of freedom (here, the `df` associated with `dose`) for the ANOVA test associated with this model.
- `df.residual` is the denominator degrees of freedom (here the `df` associated with `residual`) for that same ANOVA test.
- Remember that the F-statistic at the bottom of the summary output provides these last four statistics, as well.
- `nobs` is the number of observations (rows) used to fit the model.

Now, let's look at the remaining four summaries:

```
glance(m1) |> select(logLik:deviance) |>  
kable(digits = 1)
```

logLik	AIC	BIC	deviance
-140.1	286.3	291	5066.7

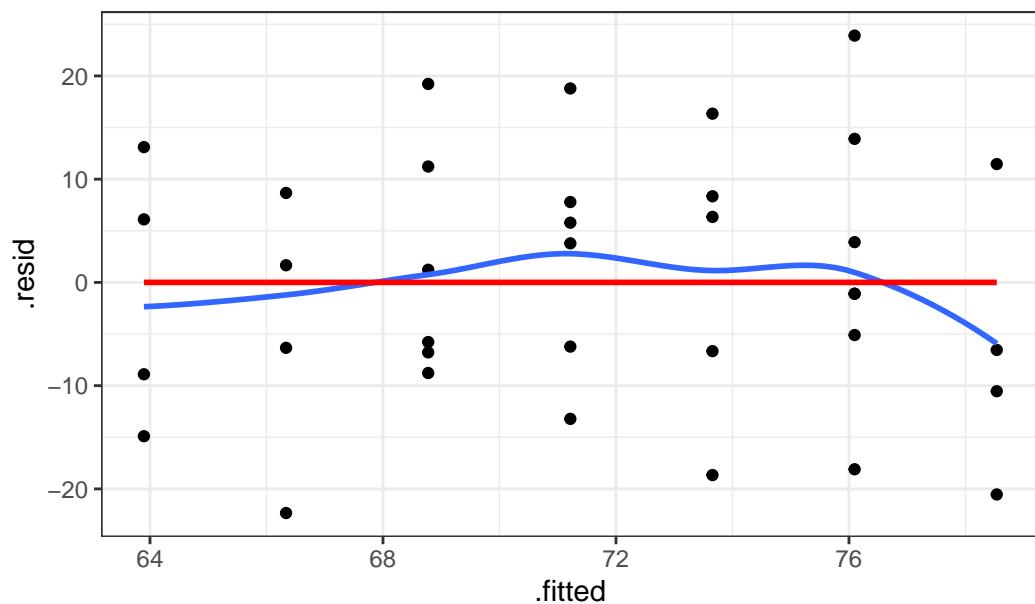
- **logLik** is the log-likelihood value for the model, and is most commonly used for a model (like the ordinary least squares model fit by `lm` that is fit using the method of maximum likelihood). Thus, the log-likelihood value will be maximized in this fit.
- **AIC** is the Akaike Information Criterion for the model. When comparing models fitted by maximum likelihood to the same outcome variable (using the same transformation, for example), the smaller the AIC, the better the fit.
- **BIC** is the Bayes Information Criterion for the model. When comparing models fitted by maximum likelihood to the same outcome variable (using the same transformation, for example), the smaller the BIC, the better the fit. BIC often prefers models with fewer coefficients to estimate than does AIC.
 - AIC and BIC can be estimated using several different approaches in R, but we'll need to use the same one across multiple models if we're comparing the results, because the concepts are only defined up to a constant.
- **deviance** is the fitted model's deviance, a measure of lack of fit. It is a generalization of the residual sum of squares seen in the ANOVA table, and takes the same value in the case of a simple linear regression model fit with `lm` as we have here. For some generalized linear models, we'll use this for hypothesis testing, just as the ANOVA table does in the linear model case.

14.12 Plotting Residuals vs. Fitted Values

To save the residuals and predicted (fitted) values from this simple regression model, we can use the `resid` and `fitted` commands, respectively, or we can use the `augment` function in the `broom` package to obtain a tidy data set containing these objects and others.

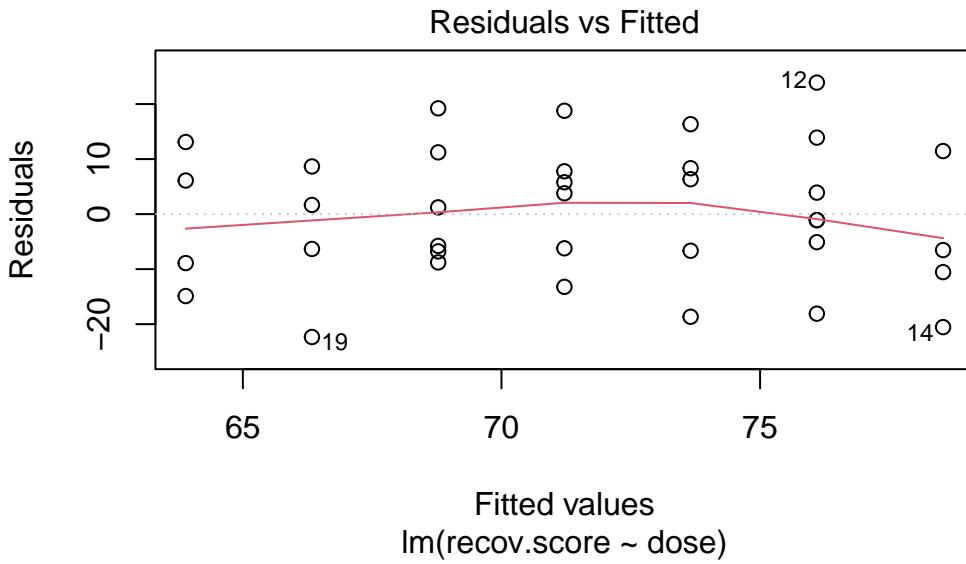
```
augment(m1) %>%  
  ggplot(., aes(x = .fitted, y = .resid)) +  
  geom_point() +  
  geom_smooth(method = "loess", formula = y ~ x, se = F) +  
  geom_smooth(method = "lm", formula = y ~ x, se = F, col = "red") +  
  labs(title = "Residuals vs. Fitted values for Model m1")
```

Residuals vs. Fitted values for Model m1



We can also obtain a plot of residuals vs. fitted values for `m1` using the following code from base R.

```
plot(m1, which = 1)
```



We hope in this plot to see a generally random scatter of points, perhaps looking like a “fuzzy football”. Since we only have seven possible `dose` values, we obtain only seven distinct predicted values, which explains the seven vertical lines in the plot. Here, the smooth red line indicates a gentle curve, but no evidence of a strong curve, or any other regular pattern in this residual plot.

15 The WCGS

15.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(knitr)
library(janitor)
library(broom)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

We will also use the `geom_density_ridges` function from the `ggridges` package, and the `favstats` function from the `mosaic` package, and the `ggpairs` function from the `GGally` package.

15.2 The Western Collaborative Group Study (wcgs) data set

Vittinghoff et al. (2012) explore data from the Western Collaborative Group Study (WCGS) in great detail¹. We'll touch lightly on some key issues in this Chapter.

The Western Collaborative Group Study (WCGS) was designed to test the hypothesis that the so-called Type A behavior pattern (TABP) - “characterized particularly by excessive drive, aggressiveness, and ambition, frequently in association with a relatively greater preoccupation with competitive activity, vocational deadlines, and similar pressures” - is a cause of coronary heart disease (CHD). Two additional goals, developed later in the study, were (1) to investigate the comparability of formulas developed in WCGS and in the Framingham Study (FS) for prediction of CHD risk, and (2) to determine how addition of TABP to an existing

¹For more on the WCGS, you might look at <http://www.epi.umn.edu/cvdepi/study-synopsis/western-collaborative-group-study/>

multivariate prediction formula affects ability to select subjects for intervention programs.

The study enrolled over 3,000 men ages 39-59 who were employed in San Francisco or Los Angeles, during 1960 and 1961.

```
wcgs <- read_csv("data/wcgs.csv") |>
  mutate(across(where(is.character), as_factor))

wcgs

# A tibble: 3,154 x 22
   id    age agec   height weight lnwght wghtcat   bmi    sbp lnsbp    dbp    chol
   <dbl> <dbl> <fct>  <dbl>   <dbl>   <dbl> <fct>   <dbl> <dbl> <dbl> <dbl> <dbl>
 1 2343    50 46-50     67    200    5.30 170-200  31.3   132   4.88    90   249
 2 3656    51 51-55     73    192    5.26 170-200  25.3   120   4.79    74   194
 3 3526    59 56-60     70    200    5.30 170-200  28.7   158   5.06    94   258
 4 22057   51 51-55     69    150    5.01 140-170  22.1   126   4.84    80   173
 5 12927   44 41-45     71    160    5.08 140-170  22.3   126   4.84    80   214
 6 16029   47 46-50     64    158    5.06 140-170  27.1   116   4.75    76   206
 7 3894    40 35-40     70    162    5.09 140-170  23.2   122   4.80    78   190
 8 11389   41 41-45     70    160    5.08 140-170  23.0   130   4.87    84   212
 9 12681   50 46-50     71    195    5.27 170-200  27.2   112   4.72    70   130
10 10005   43 41-45     68    187    5.23 170-200  28.4   120   4.79    80   233
# ... with 3,144 more rows, and 10 more variables: behpat <fct>, dibpat <fct>,
#   smoke <fct>, ncigs <dbl>, arcus <dbl>, chd69 <fct>, typchd69 <dbl>,
#   time169 <dbl>, t1 <dbl>, uni <dbl>
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Here, we have 3154 rows (subjects) and 22 columns (variables). After importing the data and creating a tibble with `read_csv`, I used `mutate(across(where(is.character), as_factor))` to convert all variables containing character data into factors.

15.2.1 Structure of `wcgs`

We can specify the (sometimes terrible) variable names, through the `names` function, or we can add other elements of the structure, so that we can identify elements of particular interest.

```
str(wcgs)
```

```

tibble [3,154 x 22] (S3: tbl_df/tbl/data.frame)
$ id      : num [1:3154] 2343 3656 3526 22057 12927 ...
$ age     : num [1:3154] 50 51 59 51 44 47 40 41 50 43 ...
$ agec    : Factor w/ 5 levels "46-50","51-55",...: 1 2 3 2 4 1 5 4 1 4 ...
$ height  : num [1:3154] 67 73 70 69 71 64 70 70 71 68 ...
$ weight  : num [1:3154] 200 192 200 150 160 158 162 160 195 187 ...
$ lnwght  : num [1:3154] 5.3 5.26 5.3 5.01 5.08 ...
$ wghtcat: Factor w/ 4 levels "170-200","140-170",...: 1 1 1 2 2 2 2 1 1 ...
$ bmi     : num [1:3154] 31.3 25.3 28.7 22.1 22.3 ...
$ sbp     : num [1:3154] 132 120 158 126 126 116 122 130 112 120 ...
$ lnsbp   : num [1:3154] 4.88 4.79 5.06 4.84 4.84 ...
$ dbp     : num [1:3154] 90 74 94 80 80 76 78 84 70 80 ...
$ chol    : num [1:3154] 249 194 258 173 214 206 190 212 130 233 ...
$ behpat  : Factor w/ 4 levels "A1","A2","B3",...: 1 1 1 1 1 1 1 1 1 1 ...
$ dibpat  : Factor w/ 2 levels "Type A","Type B": 1 1 1 1 1 1 1 1 1 1 ...
$ smoke   : Factor w/ 2 levels "Yes","No": 1 1 2 2 2 1 2 1 2 1 ...
$ ncigs   : num [1:3154] 25 25 0 0 0 80 0 25 0 25 ...
$ arcus   : num [1:3154] 1 0 1 1 0 0 0 0 1 0 ...
$ chd69   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
$ typchd69: num [1:3154] 0 0 0 0 0 0 0 0 0 0 ...
$ time169 : num [1:3154] 1367 2991 2960 3069 3081 ...
$ t1      : num [1:3154] -1.63 -4.06 0.64 1.12 2.43 ...
$ uni     : num [1:3154] 0.486 0.186 0.728 0.624 0.379 ...

```

15.2.2 Codebook for wcgs

This table was lovingly hand-crafted, and involved a lot of typing. We'll look for better ways in 432.

Name	Stored As	Type	Details (units, levels, etc.)
id	integer	(nominal)	ID #, nominal and uninteresting
age	integer	quantitative	age, in years - no decimal places
agec	factor (5)	(ordinal)	age: 35-40, 41-45, 46-50, 51-55, 56-60
height	integer	quantitative	height, in inches
weight	integer	quantitative	weight, in pounds
lnwght	number	quantitative	natural logarithm of weight
wghtcat	factor (4)	(ordinal)	wt: < 140, 140-170, 170-200, > 200
bmi	number	quantitative	body-mass index: $703 * \text{weight in lb} / (\text{height in in})^2$
sbp	integer	quantitative	systolic blood pressure, in mm Hg
lnsbp	number	quantitative	natural logarithm of sbp
dbp	integer	quantitative	diastolic blood pressure, mm Hg

Name	Stored As	Type	Details (units, levels, etc.)
chol	integer	quantitative	total cholesterol, mg/dL
behpat	factor (4)	(nominal)	behavioral pattern: A1, A2, B3 or B4
dibpat	factor (2)	(binary)	behavioral pattern: A or B
smoke	factor (2)	(binary)	cigarette smoker: Yes or No
ncigs	integer	quantitative	number of cigarettes smoked per day
arcus	integer	(nominal)	arcus senilis present (1) or absent (0)
chd69	factor (2)	(binary)	CHD event: Yes or No
typchd69	integer	(4 levels)	event: 0 = no CHD, 1 = MI or SD, 2 = silent MI, 3 = angina
time169	integer	quantitative	follow-up time in days
t1	number	quantitative	heavy-tailed (random draws)
uni	number	quantitative	light-tailed (random draws)

15.2.3 Quick Summary

```
summary(wcgs)
```

id	age	agec	height	weight
Min. : 2001	Min. :39.00	46-50: 750	Min. :60.00	Min. : 78
1st Qu.: 3741	1st Qu.:42.00	51-55: 528	1st Qu.:68.00	1st Qu.:155
Median :11406	Median :45.00	56-60: 242	Median :70.00	Median :170
Mean :10478	Mean :46.28	41-45:1091	Mean :69.78	Mean :170
3rd Qu.:13115	3rd Qu.:50.00	35-40: 543	3rd Qu.:72.00	3rd Qu.:182
Max. :22101	Max. :59.00		Max. :78.00	Max. :320
lnwght	wghtcat	bmi	sbp	lnsbp
Min. :4.357	170-200:1171	Min. :11.19	Min. : 98.0	Min. :4.585
1st Qu.:5.043	140-170:1538	1st Qu.:22.96	1st Qu.:120.0	1st Qu.:4.787
Median :5.136	> 200 : 213	Median :24.39	Median :126.0	Median :4.836
Mean :5.128	< 140 : 232	Mean :24.52	Mean :128.6	Mean :4.850
3rd Qu.:5.204		3rd Qu.:25.84	3rd Qu.:136.0	3rd Qu.:4.913
Max. :5.768		Max. :38.95	Max. :230.0	Max. :5.438
dbp	chol	behpat	dibpat	smoke
Min. : 58.00	Min. :103.0	A1: 264	Type A:1589	Yes:1502
1st Qu.: 76.00	1st Qu.:197.2	A2:1325	Type B:1565	No :1652
Median : 80.00	Median :223.0	B3:1216		
Mean : 82.02	Mean :226.4	B4: 349		
3rd Qu.: 86.00	3rd Qu.:253.0			

```

Max.    :150.00   Max.    :645.0
NA's     :12

ncigs      arcus      chd69      typchd69      time169
Min.    : 0.0    Min.    :0.0000    No :2897    Min.    :0.0000    Min.    : 18
1st Qu.: 0.0    1st Qu.:0.0000    Yes: 257   1st Qu.:0.0000    1st Qu.:2842
Median   : 0.0    Median   :0.0000          Median :0.0000    Median  :2942
Mean     :11.6    Mean     :0.2985          Mean   :0.1363    Mean    :2684
3rd Qu.:20.0    3rd Qu.:1.0000          3rd Qu.:0.0000    3rd Qu.:3037
Max.     :99.0    Max.     :1.0000          Max.   :3.0000    Max.    :3430
NA's     :2

t1          uni
Min.    :-47.43147  Min.    :0.0007097
1st Qu.: -1.00337  1st Qu.:0.2573755
Median   : 0.00748  Median   :0.5157779
Mean     : -0.03336 Mean     :0.5052159
3rd Qu.:  0.97575  3rd Qu.:0.7559902
Max.     : 47.01623 Max.     :0.9994496
NA's     :39

```

For a more detailed description, we might consider `Hmisc::describe`, `psych::describe`, `mosaic::favstats`, etc.

15.3 Are the SBPs Normally Distributed?

Consider the question of whether the distribution of the systolic blood pressure results is well-approximated by the Normal.

```
res <- mosaic::favstats(~ sbp, data = wcgs)
```

```

Registered S3 method overwritten by 'mosaic':
  method                  from
  fortify.SpatialPolygonsDataFrame ggplot2

bin_w <- 5 # specify binwidth

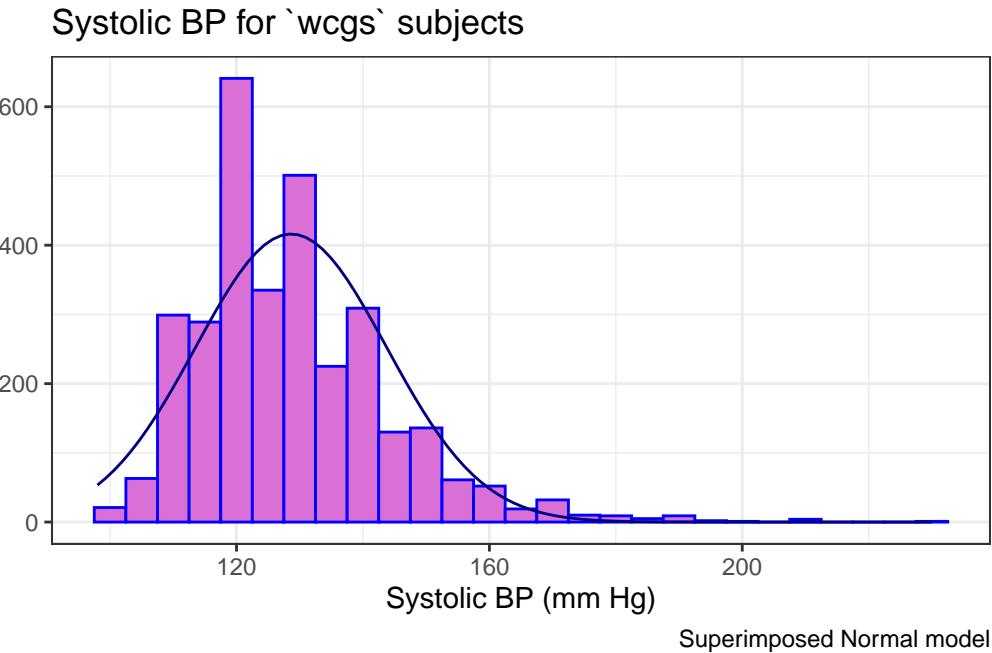
ggplot(wcgs, aes(x = sbp)) +
  geom_histogram(binwidth = bin_w,
                 fill = "orchid",
                 col = "blue") +

```

```

stat_function(
  fun = function(x) dnorm(x, mean = res$mean,
                           sd = res$sd) *
    res$n * bin_w,
  col = "navy") +
  labs(title = "Systolic BP for `wcgs` subjects",
       x = "Systolic BP (mm Hg)", y = "",
       caption = "Superimposed Normal model")

```



Since the data contain both `sbp` and `lnsbp` (its natural logarithm), let's compare them. Note that in preparing the graph, we'll need to change the location for the text annotation.

```

res <- mosaic::favstats(~ lnsbp, data = wcgs)
bin_w <- 0.05 # specify binwidth

ggplot(wcgs, aes(x = lnsbp)) +
  geom_histogram(binwidth = bin_w,
                 fill = "orange",
                 col = "blue") +
  stat_function(
    fun = function(x) dnorm(x, mean = res$mean,

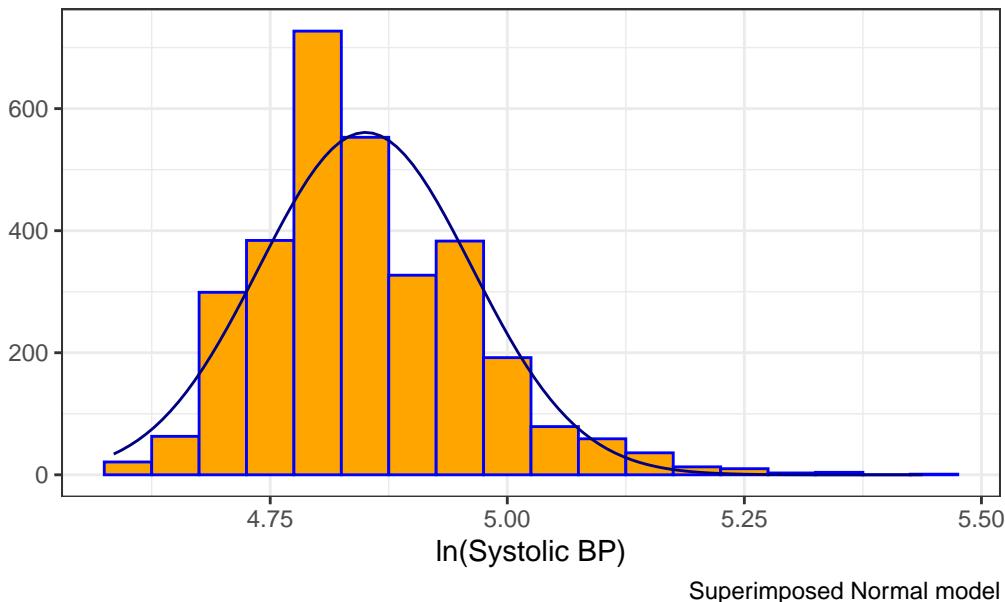
```

```

sd = res$sd) *
res$n * bin_w,
col = "navy") +
labs(title = "ln(Systolic BP) for `wcgs` subjects",
x = "ln(Systolic BP)", y = "",
caption = "Superimposed Normal model")

```

ln(Systolic BP) for `wcgs` subjects



We can also look at Normal Q-Q plots, for instance...

```

p1 <- ggplot(wcgs, aes(sample = sbp)) +
  geom_qq(color = "orchid") +
  geom_qq_line(color = "red") +
  labs(y = "Ordered SBP", title = "sbp in wcgs")

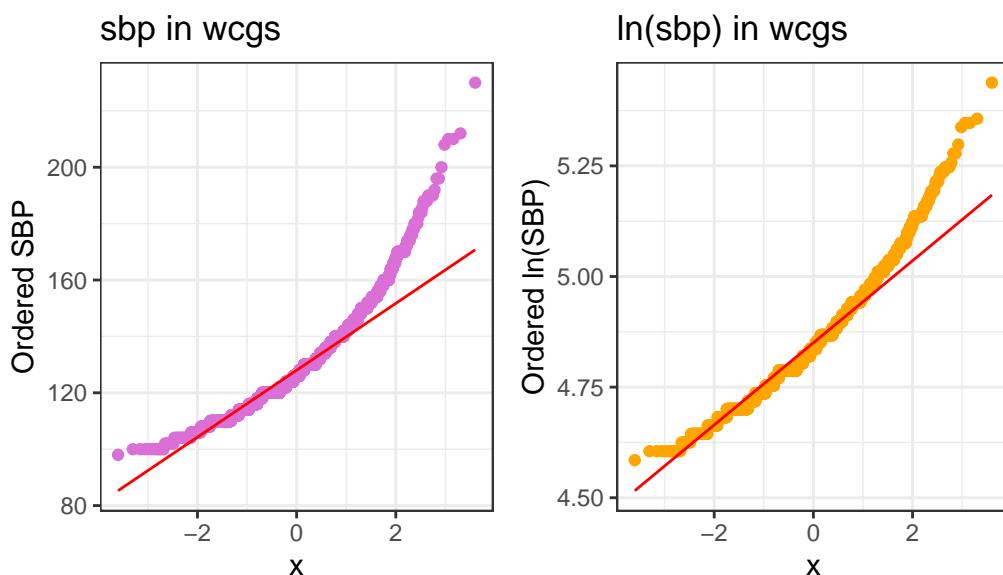
p2 <- ggplot(wcgs, aes(sample = lnsbp)) +
  geom_qq(color = "orange") +
  geom_qq_line(color = "red") +
  labs(y = "Ordered ln(SBP)", title = "ln(sbp) in wcgs")

## next step requires library(patchwork)

```

```
p1 + p2 +
  plot_annotation(title = "Normal Q-Q plots of SBP and ln(SBP) in wcgs")
```

Normal Q–Q plots of SBP and $\ln(\text{SBP})$ in `wcgs`



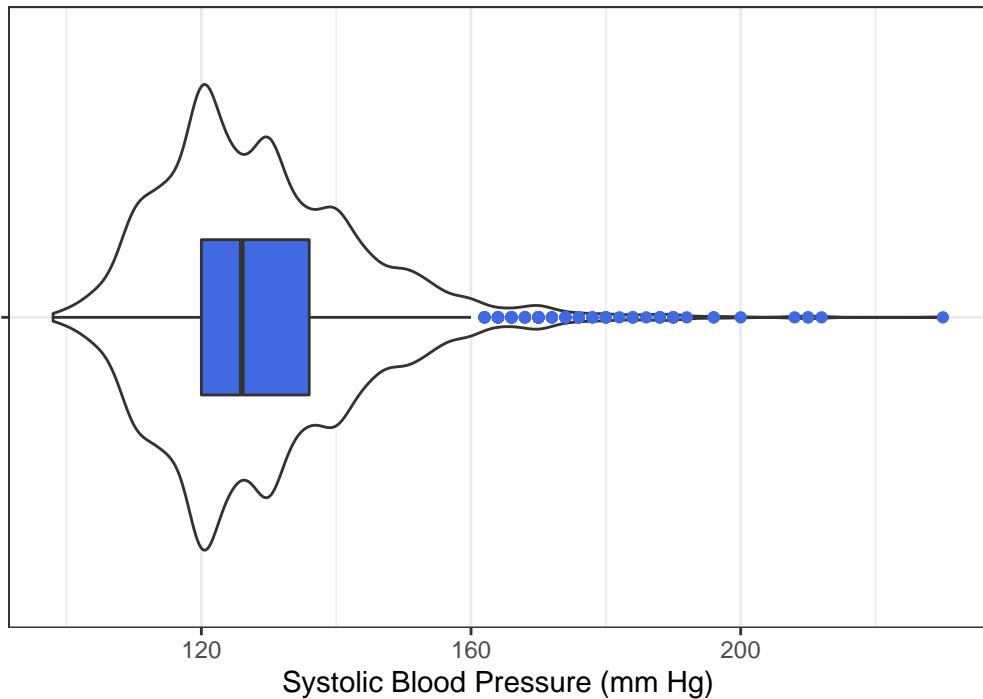
There's at best a small improvement from `sbp` to `lnsbp` in terms of approximation by a Normal distribution.

15.4 Identifying and Describing SBP outliers

It looks like there's an outlier (or a series of them) in the SBP data.

```
ggplot(wcgs, aes(x = "", y = sbp)) +
  geom_violin() +
  geom_boxplot(width = 0.3, fill = "royalblue",
               outlier.color = "royalblue") +
  labs(title = "Boxplot with Violin of SBP in `wcgs` data",
       y = "Systolic Blood Pressure (mm Hg)",
       x = "") +
  coord_flip()
```

Boxplot with Violin of SBP in `wcgs` data



```
Hmisc::describe(wcgs$sbp)
```

	n	missing	distinct	Info	Mean	Gmd	.05	.10
3154		0	62	0.996	128.6	16.25	110	112
	.25	.50	.75	.90	.95			
	120	126	136	148	156			

lowest : 98 100 102 104 106, highest: 200 208 210 212 230

The maximum value here is 230, and is clearly the most extreme value in the data set. One way to gauge this is to describe that observation's **Z score**, the number of standard deviations away from the mean that the observation falls. Here, the maximum value, 230 is 6.71 standard deviations above the mean, and thus has a Z score of 6.7.

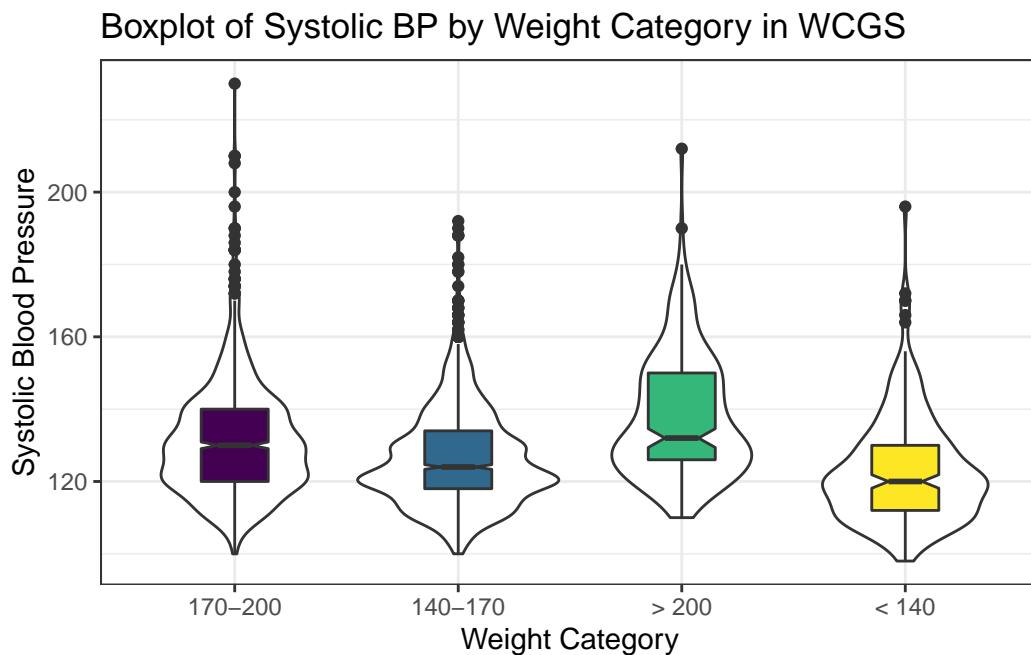
A negative Z score would indicate a point below the mean, while a positive Z score indicates, as we've seen, a point above the mean. The minimum systolic blood pressure, 98 is 2.03 standard deviations *below* the mean, so it has a Z score of -2.

Recall that the Empirical Rule suggests that if a variable follows a Normal distribution, it would have approximately 95% of its observations falling inside a Z score of (-2, 2), and 99.74% falling inside a Z score range of (-3, 3). Do the systolic blood pressures appear Normally distributed?

15.5 Does Weight Category Relate to SBP?

The data are collected into four groups based on the subject's weight (in pounds).

```
ggplot(wcgs, aes(x = wghtcat, y = sbp)) +  
  geom_violin() +  
  geom_boxplot(aes(fill = wghtcat), width = 0.3, notch = TRUE) +  
  scale_fill_viridis_d() +  
  guides(fill = "none") +  
  labs(title = "Boxplot of Systolic BP by Weight Category in WCGS",  
       x = "Weight Category", y = "Systolic Blood Pressure")
```



15.6 Re-Leveling a Factor

Well, that's not so good. We really want those weight categories (the *levels*) to be ordered more sensibly.

```
wcgs |> tabyl(wghtcat)
```

```
wghtcat      n      percent
170-200 1171 0.37127457
140-170 1538 0.48763475
> 200   213 0.06753329
< 140   232 0.07355739
```

Like all *factor* variables in R, the categories are specified as levels. We want to change the order of the levels in a new version of this factor variable so they make sense. There are multiple ways to do this, but I prefer the `fct_relevel` function from the `forcats` package (part of the tidyverse.) Which order is more appropriate?

I'll add a new variable to the `wcgs` data called `weight_f` that relevels the `wghtcat` data.

```
wcgs <- wcgs |>
  mutate(weight_f = fct_relevel(wghtcat, "< 140", "140-170", "170-200", "> 200"))

wcgs |> tabyl(weight_f)

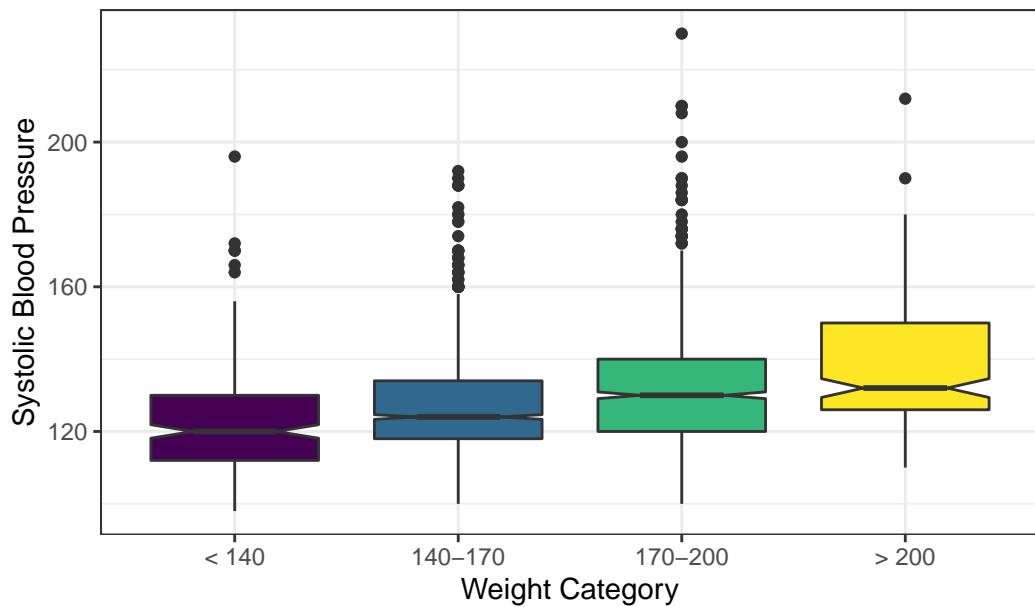
weight_f      n      percent
< 140   232 0.07355739
140-170 1538 0.48763475
170-200 1171 0.37127457
> 200   213 0.06753329
```

For more on the `forcats` package, check out Wickham and Grolemund (2022), especially the Section on Factors.

15.6.1 SBP by Weight Category

```
ggplot(wcgs, aes(x = weight_f, y = sbp, fill = weight_f)) +
  geom_boxplot(notch = TRUE) +
  scale_fill_viridis_d() +
  guides(fill = "none") +
  labs(title = "Systolic Blood Pressure by Reordered Weight Category in WCGS",
       x = "Weight Category", y = "Systolic Blood Pressure")
```

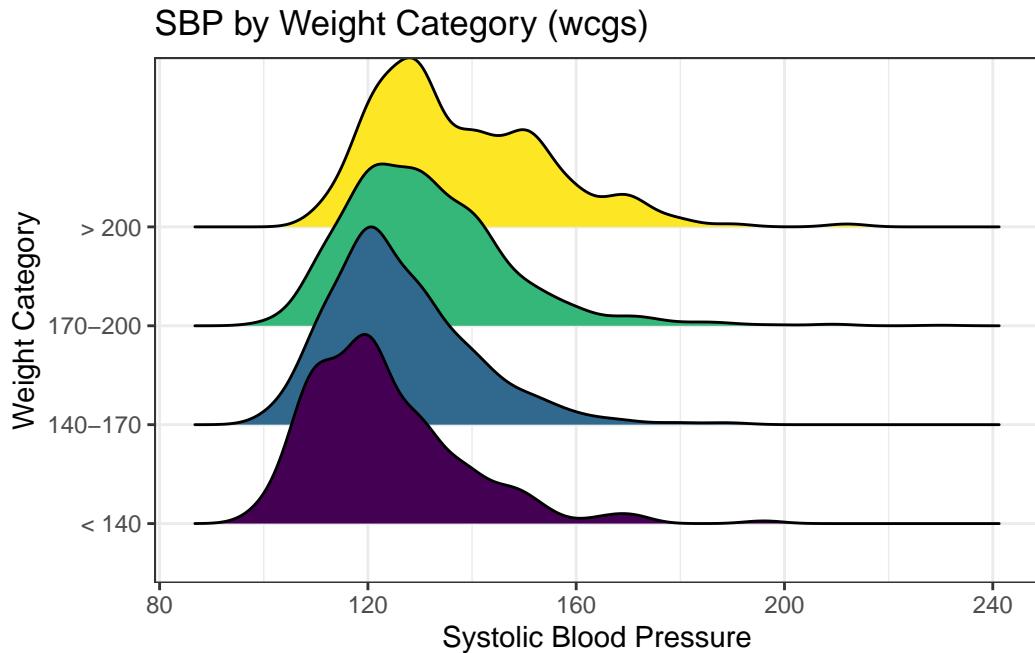
Systolic Blood Pressure by Reordered Weight Category in WC



We might see some details well with a **ridgeline plot**, too.

```
ggplot(wcgs, aes(x = sbp, y = weight_f, fill = weight_f, height = ..density..)) +  
  ggridges::geom_density_ridges(scale = 2) +  
  scale_fill_viridis_d() +  
  guides(fill = "none") +  
  labs(title = "SBP by Weight Category (wcgs)",  
       x = "Systolic Blood Pressure",  
       y = "Weight Category")
```

Picking joint bandwidth of 3.74



As the plots suggest, patients in the heavier groups generally had higher systolic blood pressures.

```
mosaic::favstats(sbp ~ weight_f, data = wcgs)
```

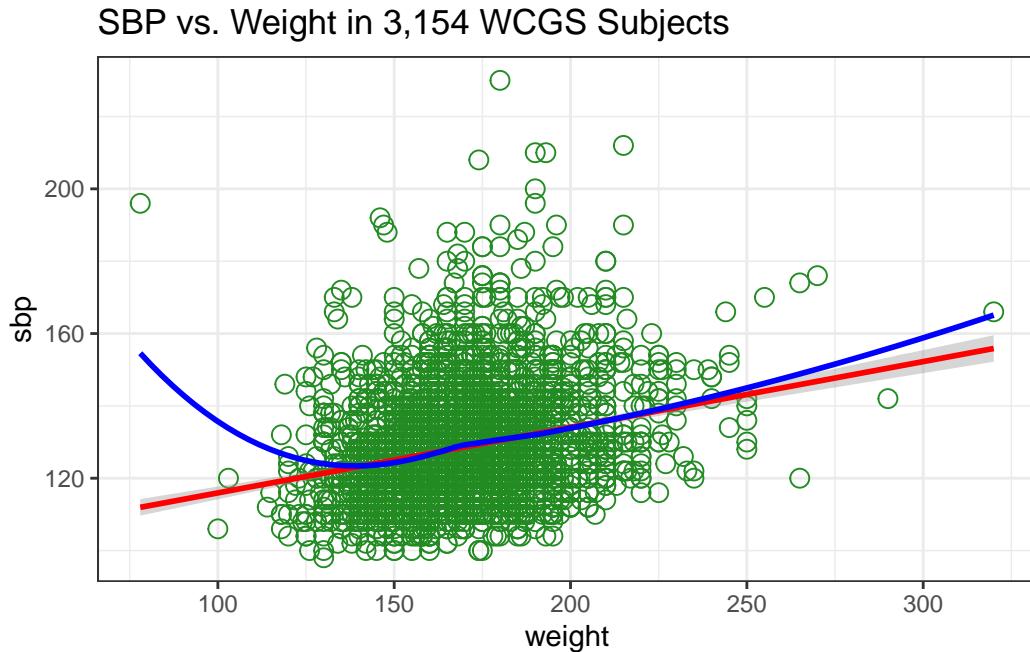
	weight_f	min	Q1	median	Q3	max	mean	sd	n	missing
1	< 140	98	112	120	130	196	123.1379	14.73394	232	0
2	140-170	100	118	124	134	192	126.2939	13.65294	1538	0
3	170-200	100	120	130	140	230	131.1136	15.57024	1171	0
4	> 200	110	126	132	150	212	137.8685	16.75522	213	0

15.7 Are Weight and SBP Linked?

Let's build a scatter plot of SBP (Outcome) by Weight (Predictor), rather than breaking down into categories.

```
ggplot(wcgs, aes(x = weight, y = sbp)) +
  geom_point(size=3, shape=1, color="forestgreen") + ## default size = 2
  stat_smooth(method=lm, color="red") + ## add se=FALSE to hide conf. interval
  stat_smooth(method=loess, se=FALSE, color="blue") +
```

```
ggttitle("SBP vs. Weight in 3,154 WCGS Subjects")
```



- The mass of the data is hidden from us - showing 3154 points in one plot can produce little more than a blur where there are lots of points on top of each other.
- Here the least squares regression line (in red), and loess scatterplot smoother, (in blue) can help.

The relationship between systolic blood pressure and weight appears to be very close to linear, but of course there is considerable scatter around that generally linear relationship. It turns out that the Pearson correlation of these two variables is 0.253.

15.8 SBP and Weight by Arcus Senilis groups?

An issue of interest to us will be to assess whether the SBP-Weight relationship we see above is similar among subjects who have arcus senilis and those who do not.

Arcus senilis is an old age syndrome where there is a white, grey, or blue opaque ring in the corneal margin (peripheral corneal opacity), or white ring in front of the periphery of the iris. It is present at birth but then fades; however, it is quite commonly present in the elderly. It can also appear earlier in life as a result of hypercholesterolemia.

Wikipedia article on Arcus Senilis, retrieved 2017-08-15

Let's start with a quick look at the `arcus` data.

```
wcgs |> tabyl(arcus)
```

arcus	n	percent	valid_percent
0	2211	0.7010145847	0.7014594
1	941	0.2983512999	0.2985406
NA	2	0.0006341154	NA

We have 2 missing values, so we probably want to do something about that before plotting the data, and we may also want to create a factor variable with more meaningful labels than 1 (which means yes, arcus senilis is present) and 0 (which means no, it isn't.)

```
wcgs <- wcgs |>  
  mutate(arcus_f = fct_recode(factor(arcus),  
    "Arcus senilis" = "1",  
    "No arcus senilis" = "0"),  
    arcus_f = fct_relevel(arcus_f, "Arcus senilis"))  
  
wcgs |> tabyl(arcus_f, arcus)
```

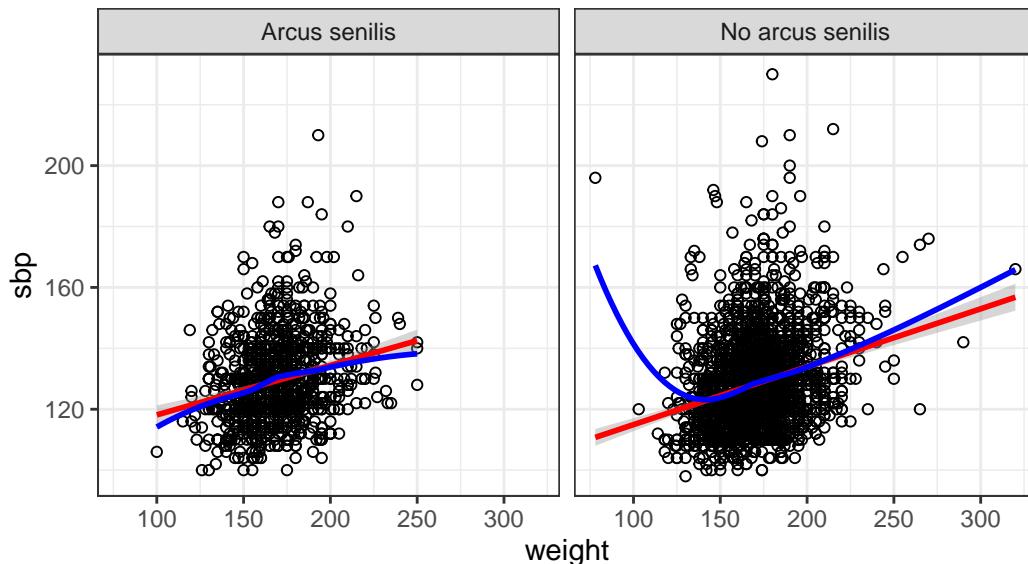
arcus_f	0	1	NA_
Arcus senilis	0	941	0
No arcus senilis	2211	0	0
<NA>	0	0	2

Let's build a version of the `wcgs` data that eliminates all missing data in the variables of immediate interest, and then plot the SBP-weight relationship in groups of patients with and without arcus senilis.

```
wcgs |>  
  filter(complete.cases(arcus_f, sbp, weight)) |>  
  ggplot(aes(x = weight, y = sbp, group = arcus_f)) +  
  geom_point(shape = 1) +  
  stat_smooth(method=lm, color="red") +  
  stat_smooth(method=loess, se=FALSE, color="blue") +  
  labs(title = "SBP vs. Weight by Arcus Senilis status",  
    caption = "3,152 Western Collaborative Group Study subjects with known arcus seni
```

```
facet_wrap(~ arcus_f)
```

SBP vs. Weight by Arcus Senilis status



15.9 Linear Model for SBP-Weight Relationship: subjects without Arcus Senilis

```
model.noarcus <-  
  lm(sbp ~ weight, data = filter(wcgs, arcus == 0))  
  
tidy(model.noarcus) |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	95.92	2.56	37.54	0
weight	0.19	0.01	12.77	0

```
glance(model.noarcus) |> select(r.squared:p.value, AIC) |> kable(digits = 3)
```

r.squared	adj.r.squared	sigma	statistic	p.value	AIC
0.069	0.068	14.799	162.959	0	18193.78

```
summary(model.noarcus)
```

Call:

```
lm(formula = sbp ~ weight, data = filter(wcgs, arcus == 0))
```

Residuals:

Min	1Q	Median	3Q	Max
-29.011	-10.251	-2.447	7.553	99.848

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	95.9219	2.5552	37.54	<2e-16 ***							
weight	0.1902	0.0149	12.77	<2e-16 ***							

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

Residual standard error: 14.8 on 2209 degrees of freedom

Multiple R-squared: 0.0687, Adjusted R-squared: 0.06828

F-statistic: 163 on 1 and 2209 DF, p-value: < 2.2e-16

The linear model for the 2211 patients without Arcus Senilis has R-squared = 6.87%.

- The regression equation is $95.92 - 0.19 \text{ weight}$, for those patients without Arcus Senilis.

15.10 Linear Model for SBP-Weight Relationship: subjects with Arcus Senilis

```
model.witharcus <-
  lm(sbp ~ weight, data = filter(wcgs, arcus == 1))

tidy(model.witharcus) |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	101.88	3.76	27.13	0
weight	0.16	0.02	7.39	0

```
glance(model.witharcus) |> select(r.squared:p.value, AIC) |> kable(digits = 3)
```

r.squared	adj.r.squared	sigma	statistic	p.value	AIC
0.055	0.054	14.192	54.583	0	7666.828

```
summary(model.witharcus)
```

Call:

```
lm(formula = sbp ~ weight, data = filter(wcgs, arcus == 1))
```

Residuals:

Min	1Q	Median	3Q	Max
-30.335	-9.636	-1.961	7.973	76.738

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	101.87847	3.75572	27.126	< 2e-16 ***
weight	0.16261	0.02201	7.388	3.29e-13 ***

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

Residual standard error: 14.19 on 939 degrees of freedom

Multiple R-squared: 0.05494, Adjusted R-squared: 0.05393

F-statistic: 54.58 on 1 and 939 DF, p-value: 3.29e-13

The linear model for the 941 patients with Arcus Senilis has R-squared = 5.49%.

- The regression equation is 101.88 - 0.163 weight, for those patients with Arcus Senilis.

15.11 Including Arcus Status in the model

```
model3 <- lm(sbp ~ weight * arcus, data = filter(wcgs, !is.na(arcus)))  
tidy(model3) |> kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	95.92	2.52	38.00	0.00
weight	0.19	0.01	12.92	0.00
arcus	5.96	4.62	1.29	0.20
weight:arcus	-0.03	0.03	-1.02	0.31

```
glance(model3) |> select(r.squared:p.value, AIC) |> kable(digits = 3)
```

r.squared	adj.r.squared	sigma	statistic	p.value	AIC
0.066	0.065	14.62	74.094	0	25860.96

```
summary(model3)
```

Call:

```
lm(formula = sbp ~ weight * arcus, data = filter(wcgs, !is.na(arcus)))
```

Residuals:

Min	1Q	Median	3Q	Max
-30.335	-10.152	-2.349	7.669	99.848

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	95.92190	2.52440	37.998	<2e-16 ***
weight	0.19017	0.01472	12.921	<2e-16 ***
arcus	5.95657	4.61972	1.289	0.197
weight:arcus	-0.02756	0.02703	-1.019	0.308

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

Residual standard error: 14.62 on 3148 degrees of freedom

Multiple R-squared: 0.06595, Adjusted R-squared: 0.06506
F-statistic: 74.09 on 3 and 3148 DF, p-value: < 2.2e-16

The actual regression equation in this setting includes both weight, and an indicator variable (1 = yes, 0 = no) for arcus senilis status, and the product term combining weight and that 1/0 indicator. In 432, we'll spend substantial time and energy discussing these product terms, but we'll not do much of that in 431.

- Note the use of the product term `weight*arcus` in the setup of the model to allow both the slope of weight and the intercept term in the model to change depending on arcus senilis status.
 - For a patient who has arcus, the regression equation is $SBP = 95.92 + 0.19 \text{ weight} + 5.96 (1) - 0.028 \text{ weight (1)} = 101.88 + 0.162 \text{ weight}$.
 - For a patient without arcus senilis, the regression equation is $SBP = 95.92 + 0.19 \text{ weight} + 5.96 (0) - 0.028 \text{ weight (0)} = 95.92 + 0.19 \text{ weight}$.

The linear model including the interaction of weight and arcus to predict sbp for the 3152 patients with known Arcus Senilis status has R-squared = 6.6%. Again, we'll discuss interaction more substantially in 432.

15.12 Predictions from these Linear Models

What is our predicted SBP for a subject weighing 175 pounds?

How does that change if our subject weighs 200 pounds?

Recall that

- *Without* Arcus Senilis, linear model for $SBP = 95.9 + 0.19 \times \text{weight}$
- *With* Arcus Senilis, linear model for $SBP = 101.9 + 0.16 \times \text{weight}$

So the predictions for a 175 pound subject are:

- $95.9 + 0.19 \times 175 = 129 \text{ mm Hg}$ without Arcus Senilis, and
- $101.9 + 0.16 \times 175 = 130 \text{ mm Hg}$ with Arcus Senilis.

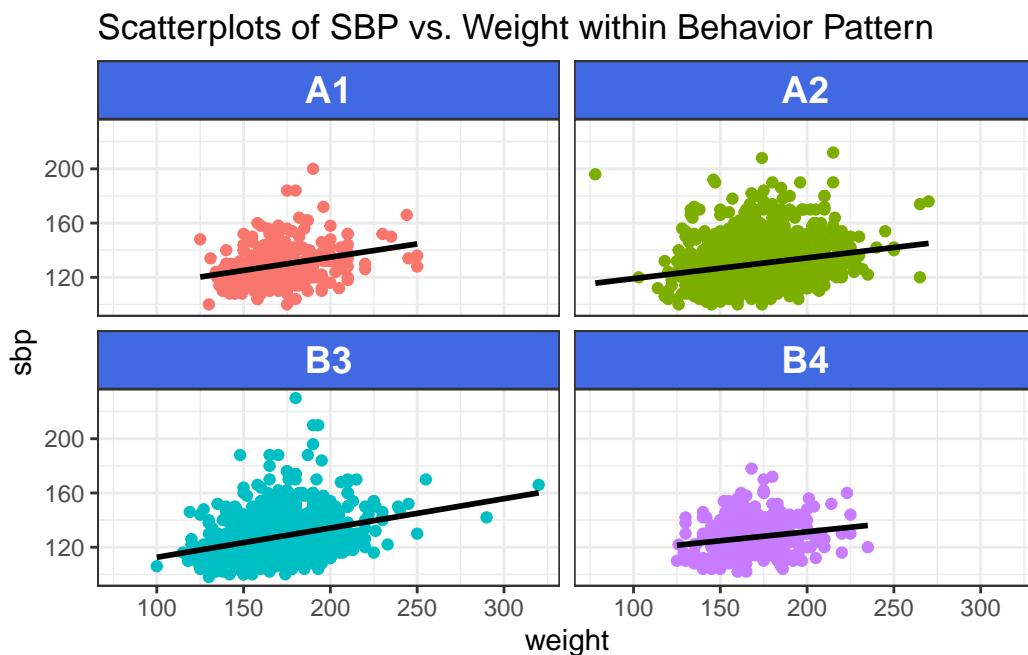
And thus, the predictions for a 200 pound subject are:

- $95.9 + 0.19 \times 200 = 134 \text{ mm Hg}$ without Arcus Senilis, and
- $101.9 + 0.16 \times 200 = 134.4 \text{ mm Hg}$ with Arcus Senilis.

15.13 Scatterplots with Facets Across a Categorical Variable

We can use facets in `ggplot2` to show scatterplots across the levels of a categorical variable, like `behpatt`.

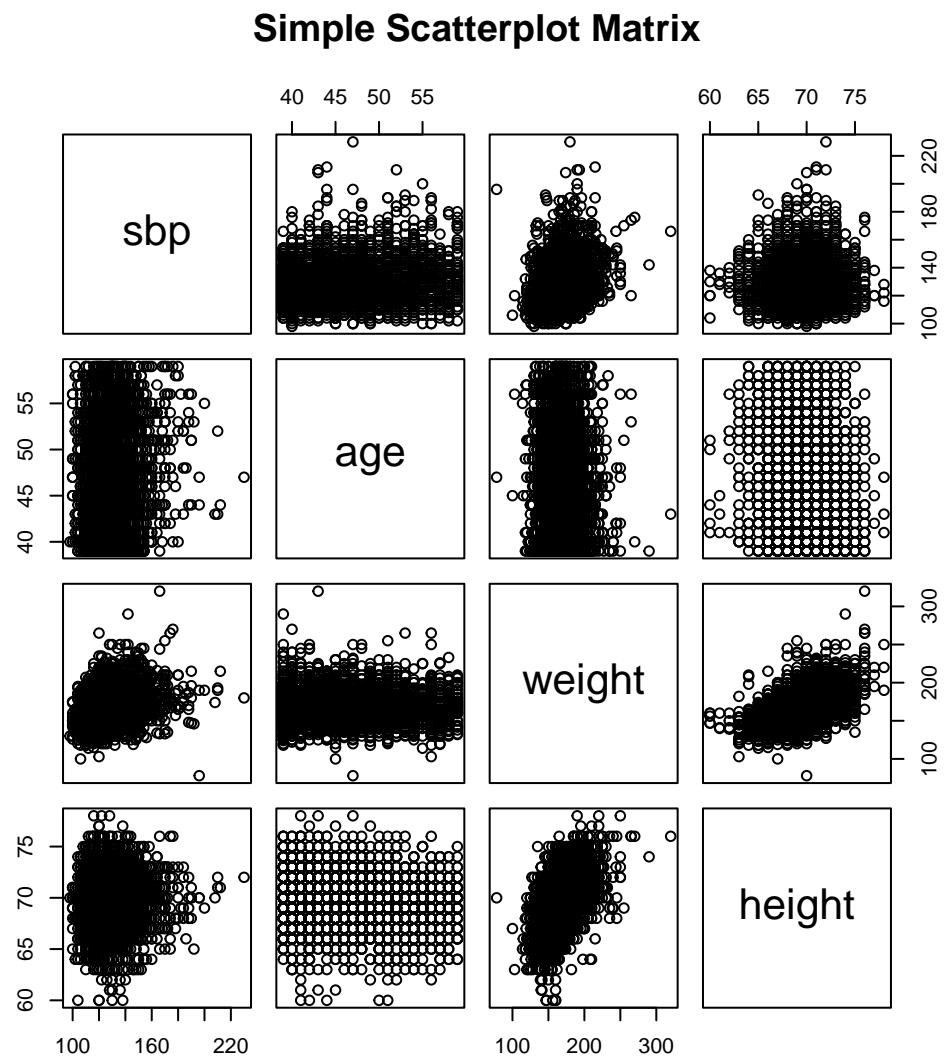
```
ggplot(wcgs, aes(x = weight, y = sbp, col = behpat)) +  
  geom_point() +  
  facet_wrap(~ behpat) +  
  geom_smooth(method = "lm", se = FALSE,  
              formula = y ~ x, col = "black") +  
  guides(color = "none") +  
  theme(strip.text = element_text(face="bold", size=rel(1.25), color="white"),  
        strip.background = element_rect(fill="royalblue")) +  
  labs(title = "Scatterplots of SBP vs. Weight within Behavior Pattern")
```



15.14 Scatterplot and Correlation Matrices

A **scatterplot matrix** can be very helpful in understanding relationships between multiple variables simultaneously. There are several ways to build such a thing, including the `pairs` function...

```
pairs (~ sbp + age + weight + height, data=wcgs, main="Simple Scatterplot Matrix")
```



15.14.1 Displaying a Correlation Matrix

```
wcgs |>  
  dplyr::select(sbp, age, weight, height) |>  
  cor() |>  
  kable(digits = 3)
```

	sbp	age	weight	height
sbp	1.000	0.166	0.253	0.018
age	0.166	1.000	-0.034	-0.095
weight	0.253	-0.034	1.000	0.533
height	0.018	-0.095	0.533	1.000

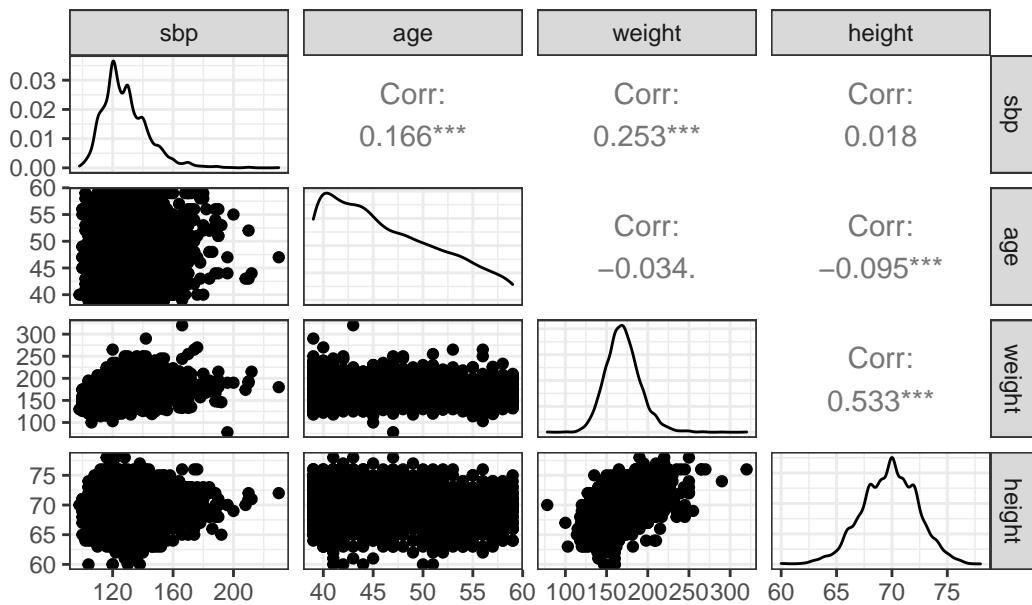
15.14.2 Using the GGally package

The `ggplot2` system doesn't have a built-in scatterplot system. There are some nice add-ins in the world, though. One option I sort of like is in the `GGally` package, which can produce both correlation matrices and scatterplot matrices.

The `ggpairs` function provides a density plot on each diagonal, Pearson correlations on the upper right and scatterplots on the lower left of the matrix.

```
GGally::ggpairs(wcgs |> select(sbp, age, weight, height),  
                 title = "Scatterplot Matrix via ggpairs")
```

Scatterplot Matrix via ggpairs



Part II

Part B. Comparing Summaries

16 Confidence Intervals for a Mean

16.1 Setup: Packages Used Here

In this part of the course, we'll make use of a few scripts I've gathered for you, in the Love-boost R script.

```
knitr::opts_chunk$set(comment = NA)

source("data/Love-boost.R")
library(broom)
library(knitr)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

We will also use the `favstats` function from the `mosaic` package, the `describe` function from the `psych` package, and several functions from the `Hmisc` package, and from the `boot` package.

16.2 Introduction

The basic theory of estimation can be used to indicate the probable accuracy and potential for bias in estimating based on limited samples. A point estimate provides a single best guess as to the value of a population or process parameter.

A confidence interval is a particularly useful way to convey to people just how much error one must allow for in a given estimate. In particular, a confidence interval allows us to quantify just how close we expect, for instance, the sample mean to be to the population or process mean. The computer will do the calculations; we need to interpret the results.

The key things that we will need to trade off are cost vs. precision, and precision vs. confidence in the correctness of the statement. Often, if we are dissatisfied with the width of the confidence interval and want to make it smaller, we have little choice but to reconsider the sample – larger samples produce shorter intervals.

16.3 This Chapter's Goals

Suppose that we are interested in learning something about a population or process, from which we can obtain a sample that consists of a subset of potential results from that population or process. The main goal for many of the parametric models that are a large part of statistics is to estimate population parameters, like a population mean, or regression coefficient, on the basis of a sample. When we do this, we want to describe not only our best guess at the parameter (referred to as a *point estimate*) but also say something useful about the uncertainty in our estimate, to let us more completely assess what the data have to tell us. A key tool for doing this is a **confidence interval**.

Essentially every textbook on introductory statistics describes the development of a confidence interval, at least for a mean. Good supplemental resources are highlighted in the references I've provided in the course syllabus.

We'll develop confidence intervals to compare parameters about two populations (either through matched pairs or independent samples) with confidence intervals soon. Here, we'll consider the problem of estimating a confidence interval to describe the mean (or median) of the population represented by a single sample of quantitative data.

16.4 Serum Zinc Levels in 462 Teenage Males (serzinc)

The `serzinc` data include serum zinc levels in micrograms per deciliter that have been gathered for a sample of 462 males aged 15-17, My source for these data is Appendix B1 of Pagano and Gauvreau (2000). Serum zinc deficiency has been associated with anemia, loss of strength and endurance, and it is thought that 25% of the world's population is at risk of zinc deficiency. Such a deficiency can indicate poor nutrition, and can affect growth and vision, for instance. "Typical" values¹ are said to be 0.66-1.10 mcg/ml, which is 66 - 110 micrograms per deciliter.

```
serzinc <- read_csv("data/serzinc.csv")  
  
Rows: 462 Columns: 2  
-- Column specification -----  
Delimiter: ","  
chr (1): ID  
dbl (1): zinc  
  
i Use `spec()` to retrieve the full column specification for this data.  
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

¹Reference values for those over the age of 10 years at <http://www.mayomedicallaboratories.com/test-catalog/Clinical+and+Interpretive/8620>, visited 2019-09-17.

```
summary(serzinc)
```

ID	zinc
Length:462	Min. : 50.00
Class :character	1st Qu.: 76.00
Mode :character	Median : 86.00
	Mean : 87.94
	3rd Qu.: 98.00
	Max. :153.00

16.5 Our Goal: A Confidence Interval for the Population Mean

After we assess the data a bit, and are satisfied that we understand it, our first inferential goal will be to produce a **confidence interval for the true (population) mean** of males age 15-17 based on this sample, assuming that these 462 males are a random sample from the population of interest, that each serum zinc level is drawn independently from an identical distribution describing that population.

To do this, we will have several different procedures available, including:

1. A confidence interval for the population mean based on a t distribution, when we assume that the data are drawn from an approximately Normal distribution, using the sample standard deviation. (Interval corresponding to a t test, and it will be a good choice when the data really are approximately Normally distributed.)
2. A resampling approach to generate a bootstrap confidence interval for the population mean, which does not require that we assume either that the population standard deviation is known, nor that the data are drawn from an approximately Normal distribution, but which has some other weaknesses.
3. A rank-based procedure called the Wilcoxon signed rank test can also be used to yield a confidence interval statement about the population pseudo-median, a measure of the population distribution's center (but not the population's mean).

16.6 Exploratory Data Analysis for Serum Zinc

16.6.1 Graphical Summaries

The code presented below builds:

- a histogram (with Normal model superimposed),
- a boxplot (with median notch) and

- a Normal Q-Q plot (with guiding straight line through the quartiles)

for the `zinc` results from the `serzinc` tibble.

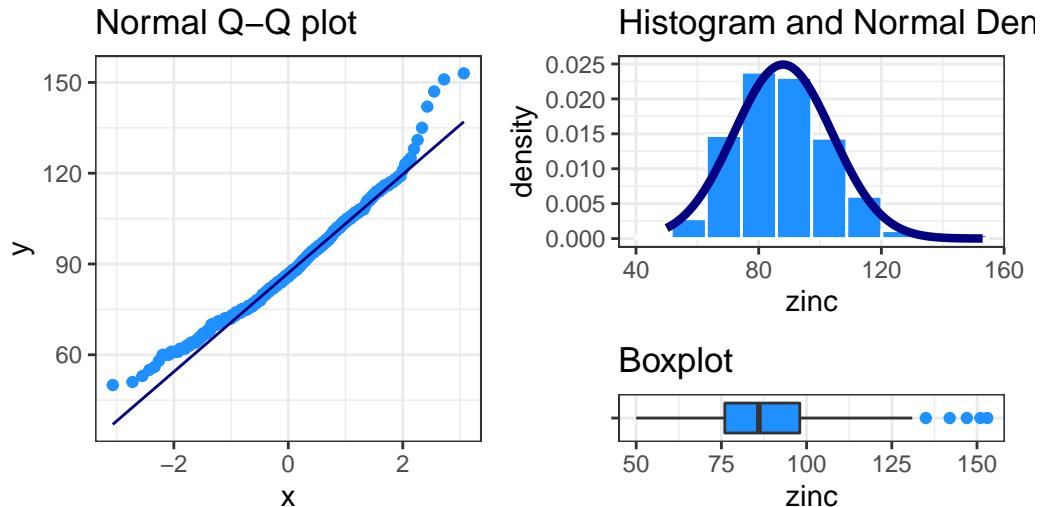
```
p1 <- ggplot(serzinc, aes(sample = zinc)) +
  geom_qq(col = "dodgerblue") + geom_qq_line(col = "navy") +
  theme(aspect.ratio = 1) +
  labs(title = "Normal Q-Q plot")

p2 <- ggplot(serzinc, aes(x = zinc)) +
  geom_histogram(aes(y = stat(density)),
                 bins = 10, fill = "dodgerblue", col = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean(serzinc$zinc),
                            sd = sd(serzinc$zinc)),
                col = "navy", lwd = 1.5) +
  labs(title = "Histogram and Normal Density")

p3 <- ggplot(serzinc, aes(x = zinc, y = "")) +
  geom_boxplot(fill = "dodgerblue", outlier.color = "dodgerblue") +
  labs(title = "Boxplot", y = "")

p1 + (p2 / p3 + plot_layout(heights = c(4,1))) +
  plot_annotation(title = "Serum Zinc (micrograms per deciliter) for 462 Teenage Males")
```

Serum Zinc (micrograms per deciliter) for 462 Teenage Males



These results include some of the more useful plots and numerical summaries when assessing shape, center and spread. The `zinc` data in the `serzinc` data frame appear to be slightly right skewed, with five outlier values on the high end of the scale, in particular.

16.6.2 Numerical Summaries

This section describes some numerical summaries of interest to augment the plots in summarizing the center, spread and shape of the distribution of serum zinc among these 462 teenage males.

```
mosaic::favstats(~ zinc, data = serzinc) |>
  kable(digits = 3)
```

	min	Q1	median	Q3	max	mean	sd	n	missing
	50	76	86	98	153	87.937	16.005	462	0

```
serzinc |>
  summarize(mean(zinc), median(zinc), sd(zinc),
            skew1 = (mean(zinc) - median(zinc))/sd(zinc)) |>
```

```
kable(digits = 3)
```

mean(zinc)	median(zinc)	sd(zinc)	skew1
87.937	86	16.005	0.121

The skew1 value here (mean - median divided by the standard deviation) backs up our graphical assessment, that the data are slightly right skewed.

```
psych::describe(serzinc$zinc)
```

```
vars   n   mean  sd median trimmed   mad min  max range skew kurtosis    se
X1     1 462 87.94 16      86  87.17 16.31  50 153   103 0.62      0.87 0.74
```

Rounded to two decimal places, the standard deviation of the serum zinc data turns out to be 16, and so the standard error of the mean, shown as `se` in the `psych::describe` output, is 16 divided by the square root of the sample size, $n = 462$. This standard error is about to become quite important to us in building statistical inferences about the mean of the entire population of teenage males based on this sample.

16.7 Defining a Confidence Interval

A confidence interval for a population or process mean uses data from a sample (and perhaps some additional information) to identify a range of potential values for the population mean, which, if certain assumptions hold, can be assumed to provide a reasonable estimate for the true population mean. A confidence interval consists of:

1. An interval estimate describing the population parameter of interest (here the population mean), and
2. A probability statement, expressed in terms of a confidence level.

16.8 Estimating the Population Mean from the Serum Zinc data

As an example, suppose that we are willing to assume that the mean serum zinc level across the entire population of teenage males, μ , follows a Normal distribution (and so, summarizing it with a mean is a rational thing to do.) Suppose that we are also willing to assume that the 462 teenage males contained in the `serzinc` tibble are a random sample from that complete

population. While we know the mean of the sample of 462 boys, we don't know μ , the mean across all teenage males. So we need to estimate it.

Earlier we estimated that a 90% confidence interval for the mean serum zinc level (μ) across the entire population of teenage males was (86.71, 89.16) micrograms per deciliter. How should we interpret this result?

- Some people think this means that there is a 90% chance that the true mean of the population, μ , falls between 86.71 and 89.16 micrograms per deciliter. That's not correct.
- The population mean is a constant **parameter** of the population of interest. That constant is not a random variable, and does not change. So the actual probability of the population mean falling inside that range is either 0 or 1.
- Our confidence is in our process.
 - It's in the sampling method (random sampling) used to generate the data, and in the assumption that the population follows a Normal distribution.
 - It's captured in our accounting for one particular type of error (called *sampling error*) in developing our interval estimate, while assuming all other potential sources of error are negligible.

So, what's closer to the truth is:

- If we used this same method to sample data from the true population of teenage males, and built 100 such 90% confidence intervals, then about 90 of them would contain the true population mean.

16.9 Confidence vs. Significance Level

We've estimated a 90% confidence interval for the population mean serum zinc level among teenage boys using the `serzinc` data.

- We call $100(1-\alpha)\%$, here, 90%, or 0.90, the *confidence* level, and
- $\alpha = 10\%$, or 0.10 is called the *significance* level.

If we had instead built a series of 100 different 95% confidence intervals, then about 95 of them would contain the true value of μ .

Let's look more closely at the issue of estimating a population **mean** based on a sample of observations. We will need three critical pieces - the sample, the confidence level, and the margin of error, which is based on the standard error of a sample mean, when we are estimating a population mean.

16.10 The Standard Error of a Sample Mean

The standard error, generally, is the name we give to the standard deviation associated with any particular parameter estimate.

- If we are using a sample mean based on a sample of size n to estimate a population mean, the **standard error of that sample mean** is the standard deviation of the measurements in the population, divided by the square root of the sample size.
- We often estimate this particular standard error with s (the sample standard deviation) divided by the square root of the sample size.
- Other statistics have different standard errors.
 - $\sqrt{p(1-p)/n}$ is the standard error of the sample proportion p estimated using a sample of size n .
 - $\sqrt{\frac{1-r^2}{n-2}}$ is the standard error of the sample Pearson correlation r estimated using n pairs of observations.
 - $\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$ is the standard error of the difference between two means \bar{x}_1 and \bar{x}_2 , estimated using samples of sizes n_1 and n_2 with sample standard deviations SD_1 and SD_2 , respectively.

In developing a confidence interval for a population mean, we may be willing to assume that the data in our sample are drawn from a Normally distributed population. If so, the most common and useful means of building a confidence interval makes use of the t distribution (sometimes called Student's t) and the notion of a *standard error*.

16.11 The t distribution and CIs for a Mean

In practical settings, we will use the t distribution to estimate a confidence interval from a population mean whenever we:

- are willing to assume that the sample is drawn at random from a population or process with a Normal distribution,
- are using our sample to estimate both the mean and standard deviation, and
- have a small sample size.

16.11.1 The Formula

The two-sided $100(1 - \alpha)\%$ confidence interval (based on a t test) is:

$$\bar{x} \pm t_{\alpha/2, n-1}(s/\sqrt{n})$$

where $t_{\alpha/2, n-1}$ is the value that cuts off the top $\alpha/2$ percent of the t distribution, with $n - 1$ degrees of freedom.

We obtain the relevant cutoff value in R by substituting in values for `alphaover2` and `n-1` into the following line of R code:

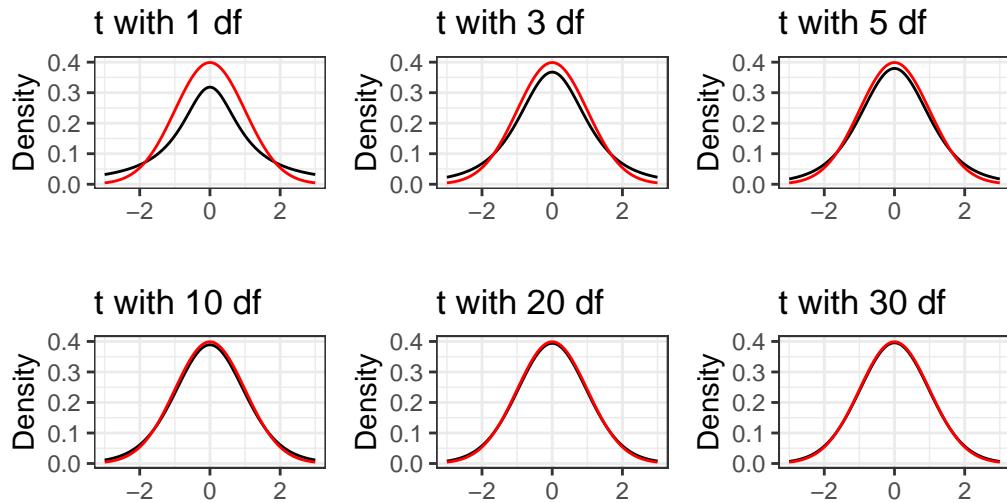
```
qt(alphaover2, df = n-1, lower.tail=FALSE)
```

16.11.2 Student's t distribution

Student's t distribution looks a lot like a Normal distribution, when the sample size is large. Unlike the normal distribution, which is specified by two parameters, the mean and the standard deviation, the t distribution is specified by one parameter, the degrees of freedom.

- t distributions with large numbers of degrees of freedom are more or less indistinguishable from the standard Normal distribution.
- t distributions with smaller degrees of freedom (say, with $df < 30$, in particular) are still symmetric, but are more outlier-prone than a Normal distribution

Various t distributions and the Standard Normal



In each plot, the Standard Normal distribution is in red

16.12 Building the CI in R

Suppose we wish to build a 90% confidence interval for the true mean serum zinc level across the entire population of teenage males. The confidence level will be 90%, or 0.90, and so the α value, which is $1 - \text{confidence} = 0.10$.

So what we know going in is that:

- We want $\alpha = 0.10$, because we're creating a 90% confidence interval.
- The sample size $n = 462$ serum zinc measurements.
- The sample mean of those measurements, $\bar{x} = 87.937$ micrograms per deciliter.
- The sample standard deviation of those measurements, $s = 16.005$ micrograms per deciliter.

16.13 Using an intercept-only regression model

in the context of fitting an intercept-only linear regression model. An intercept-only model is fitted by putting the number 1 on the right hand side of our linear model. The resulting model simply fits the overall mean of the data as a prediction for all subjects.

```

model_zinc <- lm(zinc ~ 1, data = serzinc)

summary(model_zinc)

Call:
lm(formula = zinc ~ 1, data = serzinc)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.937 -11.937  -1.937  10.063  65.063 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 87.9372    0.7446   118.1   <2e-16 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16 on 461 degrees of freedom

```

```
confint(model_zinc, level = 0.90)
```

	5 %	95 %
(Intercept)	86.71	89.16446

Generally, though, I'll use the `tidy()` function in `broom` to obtain the key information from a model like this:

```
tidy(model_zinc, conf.int = TRUE, conf = 0.90) |>
  kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	87.94	0.74	118.1	0	86.71	89.16

As an alternative, we could also use the `t.test` function, which can build (in this case) a two-sided confidence interval for the zinc levels like this:

```
tt <- t.test(serzinc$zinc,
             conf.level = 0.90,
             alternative = "two.sided")
```

```
tt
```

One Sample t-test

```
data: serzinc$zinc
t = 118.1, df = 461, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 86.71000 89.16446
sample estimates:
mean of x
87.93723
```

and the `tidy()` function from the `broom` package works here, too.

```
# requires library(broom)
tidy(tt, conf.int = TRUE, conf = 0.90) |>
  knitr::kable(digits = 2)
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
87.94	118.1	0	461	86.71	89.16	One Sample t-test	two.sided

And again, our 90% confidence interval for the true population mean serum zinc level, based on our sample of 462 patients, is (86.71, 89.16) micrograms per deciliter².

16.14 Interpreting the Result

An appropriate interpretation of the 90% two-sided confidence interval above follows:

- (86.71, 89.16) micrograms per deciliter is a 90% two-sided confidence interval for the population mean serum zinc level among teenage males.

²Since the measured zinc levels appear as integers, we should probably be rounding even further in our confidence interval, down to perhaps one decimal place.

- Our point estimate for the true population mean serum zinc level is 87.94. The values in the interval (86.71, 89.16) represent a reasonable range of estimates for the true population mean serum zinc level, and we are 90% confident that this method of creating a confidence interval will produce a result containing the true population mean serum zinc level.
- Were we to draw 100 samples of size 462 from the population described by this sample, and use each such sample to produce a confidence interval in this manner, approximately 90 of those confidence intervals would cover the true population mean serum zinc level.

16.15 What if we want a 95% or 99% confidence interval instead?

We can obtain them using `tidy` and the same modeling approach.

```
tidy(model_zinc, conf.int = TRUE, conf.level = 0.95)

# A tibble: 1 x 7
  term      estimate std.error statistic p.value conf.low conf.high
  <chr>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>     <dbl>
1 (Intercept) 87.9      0.745     118.       0     86.5     89.4

tidy(model_zinc, conf.int = TRUE, conf.level = 0.99)

# A tibble: 1 x 7
  term      estimate std.error statistic p.value conf.low conf.high
  <chr>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>     <dbl>
1 (Intercept) 87.9      0.745     118.       0     86.0     89.9
```

16.16 Using the broom package with the t test

The `broom` package takes the messy output of built-in functions in R, such as `lm`, `t.test` or `wilcox.test`, and turns them into tidy data frames. A detailed description of the package and three of its key functions is found at <https://github.com/tidyverse/broom>.

For example, we can use the `tidy` function within `broom` to create a single-row tibble of the key results from a t test.

```
tt <- t.test(serzinc$zinc, conf.level = 0.95, alternative = "two.sided")
tidy(tt)
```

```

# A tibble: 1 x 8
  estimate statistic p.value parameter conf.low conf.high method      alter~1
    <dbl>      <dbl>     <dbl>      <dbl>      <dbl>   <chr>      <chr>
1     87.9      118.       0       461      86.5      89.4 One Sample t-- two.si~
# ... with abbreviated variable name 1: alternative

```

We can thus pull the endpoints of a 95% confidence interval directly from this output. `broom` also has a `glance` function, which returns the same information as `tidy` in the case of a t-test.

16.16.1 Effect of Changing the Confidence Level

Below, we see two-sided confidence intervals for various levels of α .

Two-Sided Interval Estimate for Zinc Level Population			
Confidence Level	α	Mean, μ	Point Estimate of μ
80% or 0.80	0.20	(87, 88.9)	87.9
90% or 0.90	0.10	(86.7, 89.2)	87.9
95% or 0.95	0.05	(86.5, 89.4)	87.9
99% or 0.99	0.01	(86, 89.9)	87.9

What happens to the width of the confidence interval in this table as the confidence level changes?

16.17 One-sided vs. Two-sided Confidence Intervals

Occasionally, we want to estimate either an upper limit for the population mean μ , or a lower limit for μ , but not both.

```
t.test(serzinc$zinc, conf.level = 0.90, alternative = "greater")
```

```

One Sample t-test

data: serzinc$zinc
t = 118.1, df = 461, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0

```

```

90 percent confidence interval:
86.98161      Inf
sample estimates:
mean of x
87.93723

```

```
t.test(serzinc$zinc, conf.level = 0.90, alternative = "less")
```

One Sample t-test

```

data: serzinc$zinc
t = 118.1, df = 461, p-value = 1
alternative hypothesis: true mean is less than 0
90 percent confidence interval:
-Inf 88.89285
sample estimates:
mean of x
87.93723

```

Note the relationship between the *two-sided* 80% confidence interval, and the *one-sided* 90% confidence intervals.

Confidence	α	Type of Interval	Interval Estimate for Zinc Level
			Population Mean, μ
80% (.80)	0.20	Two-Sided	(86.98, 88.89)
90% (.90)	0.10	One-Sided (Less Than)	$\mu < 88.89$.
90% (.90)	0.10	One-Sided (Greater Than)	$\mu > 86.98$.

Why does this happen? The 80% two-sided interval is placed so as to cut off the top 10% of the distribution with its upper bound, and the bottom 10% of the distribution with its lower bound. The 90% “less than” one-sided interval is placed so as to have its lower bound cut off the top 10% of the distribution.

The same issue appears when we consider two-sided 90% and one-sided 95% confidence intervals.

Confidence	α	Type of Interval	Interval Estimate for Zinc Level
			Population Mean, μ
90% (.90)	0.10	Two-Sided	(86.71, 89.16)

Confidence	α	Type of Interval	Interval Estimate for Zinc Level
			Population Mean, μ
95% (.95)	0.05	One-Sided (Less Than)	$\mu < 89.16.$
95% (.95)	0.05	One-Sided (Greater Than)	$\mu > 86.71.$

Again, the 90% two-sided interval cuts off the top 5% and bottom 5% of the distribution with its bounds. The 95% “less than” one-sided interval also has its lower bound cut off the top 5% of the distribution.

16.18 Bootstrap Confidence Intervals

The bootstrap (and in particular, what’s known as bootstrap resampling) is a really good idea that you should know a little bit about.

If we want to know how accurately a sample mean estimates the population mean, we would ideally like to take a very, very large sample, because if we did so, we could conclude with something that would eventually approach mathematical certainty that the sample mean would be very close to the population mean.

But we can rarely draw enormous samples. So what can we do?

16.19 Resampling is A Big Idea

One way to find out how precise our estimates are is to run them on multiple samples of the same size. This *resampling* approach was codified originally by Brad Efron in 1979.

Oversimplifying a lot, the idea is that if we sample (with replacement) from our current sample, we can draw a new sample of the same size as our original.

- And if we repeat this many times, we can generate as many samples of, say, 462 zinc levels, as we like.
- Then we take these thousands of samples and calculate (for instance) the sample mean for each, and plot a histogram of those means.
- If we then cut off the top and bottom 5% of these sample means, we obtain a reasonable 90% confidence interval for the population mean.

16.20 When is a Bootstrap Confidence Interval Reasonable?

A bootstrapped interval estimate for the population mean, μ , will be reasonable as long as we're willing to believe that:

- the original sample was a random sample (or at least a completely representative sample) from a population,
- and that the samples are independent of each other,

even if the population of interest doesn't follow a Normal, or even a symmetric distribution.

A downside of the bootstrap is that you and I will get (somewhat) different answers if we resample from the same data without setting the same random seed.

16.21 Bootstrap confidence interval for the mean: Process

To avoid the Normality assumption, and take advantage of modern computing power, we use R to obtain a bootstrap confidence interval for the population mean based on a sample.

What the computer does:

1. Re-sample the data with replacement, until it obtains a new sample that is equal in size to the original data set.
2. Calculates the statistic of interest (here, a sample mean.)
3. Repeat the steps above many times (the default is 1,000 using our approach) to obtain a set of 1,000 sample means.
4. Sort those 1,000 sample means in order, and estimate the 95% confidence interval for the population mean based on the middle 95% of the 1,000 bootstrap samples.
5. Send us a result, containing the sample mean, and a 95% confidence interval for the population mean

16.22 Using R to estimate a bootstrap CI

The command that we use to obtain a Confidence Interval for μ using the basic nonparametric bootstrap and without assuming a Normally distributed population, is `smean.cl.boot`, a part of the `Hmisc` package in R.

```
set.seed(431)
Hmisc::smean.cl.boot(serzinc$zinc, B = 1000, conf.int = 0.90)
```

	Mean	Lower	Upper
87.93723	86.76775	89.20617	

- Remember that the t-based 90% CI for μ was (86.71, 89.16), according to the following output...

```
tidy(lm(zinc ~ 1, data = serzinc), conf.int = TRUE, conf.level = 0.90)
```

```
# A tibble: 1 x 7
  term      estimate std.error statistic p.value conf.low conf.high
  <chr>      <dbl>     <dbl>     <dbl>    <dbl>     <dbl>     <dbl>
1 (Intercept) 87.9      0.745     118.      0       86.7      89.2
```

16.23 Comparing Bootstrap and T-Based Confidence Intervals

- The `smean.cl.boot` function (unlike most R functions) deletes missing data automatically, as does the `smean.cl.normal` function, which can also be used to produce the t-based confidence interval.

```
set.seed(431)
Hmisc::smean.cl.boot(serzinc$zinc, B = 1000, conf.int = 0.90)
```

	Mean	Lower	Upper
87.93723	86.76775	89.20617	

```
Hmisc::smean.cl.normal(serzinc$zinc, conf.int = 0.90)
```

	Mean	Lower	Upper
87.93723	86.71000	89.16446	

Bootstrap resampling confidence intervals do not follow the general confidence interval strategy using a point estimate plus or minus a margin for error.

- A bootstrap interval is often asymmetric, and while it will generally have the point estimate (the sample mean) near its center, for highly skewed data, this will not necessarily be the case.
- We will usually use either 1,000 (the default) or 10,000 bootstrap replications for building confidence intervals – practically, it makes little difference.

16.23.1 Bootstrap Resampling: Advantages and Caveats

The bootstrap may seem like the solution to all estimation problems. In theory, we could use the same approach to find a confidence interval for any other parameter – it's not perfect, but it is very useful. Bootstrap procedures exist for virtually any statistical comparison - the t-test analog is just one of many possibilities, and bootstrap methods are rapidly gaining on more traditional approaches in the literature thanks mostly to faster computers.

The great advantage of the bootstrap is its relative simplicity, but don't forget that many of the original assumptions of the t-based confidence interval still hold.

- Using a bootstrap does eliminate the need to worry about the Normality assumption in small sample size settings, but it still requires independent and identically distributed samples from the population of interest.

The bootstrap produces clean and robust inferences (such as confidence intervals) in many tricky situations. It is still possible that the results can be both:

- **inaccurate** (i.e. they can include the true value of the unknown population mean less often than the stated confidence probability) and
- **imprecise** (i.e., they can include more extraneous values of the unknown population mean than is desirable).

16.24 Using the Bootstrap to develop other CIs

16.24.1 Changing the Confidence Level

What if we wanted to change the confidence level?

```
set.seed(431654)
Hmisc::smean.cl.boot(serzinc$zinc, B = 1000, conf.int = 0.95)
```

Mean	Lower	Upper
87.93723	86.51066	89.42002

```
set.seed(431321)
Hmisc::smean.cl.boot(serzinc$zinc, B = 1000, conf.int = 0.99)
```

Mean	Lower	Upper
87.93723	86.20657	89.68619

16.25 One-Tailed Bootstrap Confidence Intervals

If you want to estimate a one tailed confidence interval for the population mean using the bootstrap, then the procedure is as follows:

1. Determine α , the significance level you want to use in your one-sided confidence interval. Remember that α is 1 minus the confidence level. Let's assume we want a 90% one-sided interval, so $\alpha = 0.10$.
2. Double α to determine the significance level we will use in the next step to fit a two-sided confidence interval.
3. Fit a two-sided confidence interval with confidence level $100(1 - 2 * \alpha)$. Let the bounds of this interval be (a, b) .
4. The one-sided (greater than) confidence interval will have a as its lower bound.
5. The one-sided (less than) confidence interval will have b as its upper bound.

Suppose that we want to find a 95% one-sided upper bound for the population mean serum zinc level among teenage males, μ , using the bootstrap.

Since we want a 95% confidence interval, we have $\alpha = 0.05$. We double that to get $\alpha = 0.10$, which implies we need to instead fit a two-sided 90% confidence interval.

```
set.seed(43101)
Hmisc::smean.cl.boot(serzinc$zinc, B = 1000, conf.int = 0.90)
```

Mean	Lower	Upper
87.93723	86.70509	89.11266

The upper bound of this two-sided 90% CI will also be the upper bound for a 95% one-sided CI.

16.25.1 Bootstrap CI for the Population Median

If we are willing to do a small amount of programming work in R, we can obtain bootstrap confidence intervals for other population parameters besides the mean. One statistic of common interest is the median. How do we find a confidence interval for the population median using a bootstrap approach? The easiest way I know of makes use of the `boot` package, as follows.

In step 1, we specify a new function to capture the medians from our sample.

```
f.median <- function(y, id)
{   median (y[id]) }
```

In step 2, we summon the `boot` package and call the `boot.ci` function.

```
set.seed(431787)
boot::boot.ci(boot::boot (serzinc$zinc, f.median, 1000), conf=0.90, type="basic")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
```

```
CALL :
boot::boot.ci(boot.out = boot::boot(serzinc$zinc, f.median, 1000),
conf = 0.9, type = "basic")
```

```
Intervals :
Level      Basic
90%   (84, 87 )
Calculations and Intervals on Original Scale
```

This yields a 90% confidence interval for the population median serum zinc level. Recall that the sample median for the serum zinc levels in our sample of 462 teenage males was 86 micrograms per deciliter.

```
mosaic::favstats(~ zinc, data = serzinc)
```

```
min Q1 median Q3 max      mean        sd    n missing
50 76     86 98 153 87.93723 16.00469 462       0
```

Actually, the `boot.ci` function can provide up to five different types of confidence interval (see the help file) if we change to `type="all"`, and some of those other versions have attractive properties. However, we'll stick with the basic approach in 431.

16.25.2 Bootstrap CI for the IQR

If for some reason, we want to find a 95% confidence interval for the population value of the inter-quartile range via the bootstrap, we can do it.

```
IQR(serzinc$zinc)
```

```
[1] 22
```

```

f.IQR <- function(y, id)
{   IQR (y[id]) }

set.seed(431207)
boot::boot.ci(boot::boot (serzinc$zinc, f.IQR,
    conf=0.95, type="basic")

```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot::boot.ci(boot.out = boot::boot(serzinc\$zinc, f.IQR, 1000),
conf = 0.95, type = "basic")

Intervals :
Level Basic
95% (20.00, 24.24)
Calculations and Intervals on Original Scale

16.26 Wilcoxon Signed Rank Procedure for CIs

It turns out to be difficult, without the bootstrap, to estimate an appropriate confidence interval for the median of a population, which might be an appealing thing to do, particularly if the sample data are clearly not Normally distributed, so that a median seems like a better summary of the center of the data. Bootstrap procedures are available to perform the task.

The Wilcoxon signed rank approach can be used as an alternative to t-based procedures to build interval estimates for the population *pseudo-median* when the population cannot be assumed to follow a Normal distribution.

As it turns out, if you're willing to assume the population is **symmetric** (but not necessarily Normally distributed) then the pseudo-median is actually equal to the population median.

16.26.1 What is a Pseudo-Median?

The pseudo-median of a particular distribution G is the median of the distribution of $(u + v)/2$, where both u and v have the same distribution (G).

- If the distribution G is symmetric, then the pseudomedian is equal to the median.
- If the distribution is skewed, then the pseudomedian is not the same as the median.

- For any sample, the pseudomedian is defined as the median of all of the midpoints of pairs of observations in the sample.

16.27 Wilcoxon Signed Rank-based CI in R

```
wilcox.test(serzinc$zinc, conf.int = TRUE, conf.level = 0.95)
```

```
Wilcoxon signed rank test with continuity correction

data: serzinc$zinc
V = 106953, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
85.99997 88.50002
sample estimates:
(pseudo)median
87.49996
```

16.27.1 Interpreting the Wilcoxon CI for the Population Median

If we're willing to believe the `zinc` levels come from a population with a symmetric distribution, the 95% Confidence Interval for the population median would be (86, 88.5)

For a non-symmetric population, this only applies to the *pseudo-median*.

Note that the pseudo-median (87.5) is actually closer here to the sample mean (87.9) than it is to the sample median (86).

16.27.2 Using the broom package with the Wilcoxon test

We can also use the `tidy` function within `broom` to create a single-row tibble of the key results from a Wilcoxon test, so long as we run `wilcox.test` specifying that we want a confidence interval.

```
wt <- wilcox.test(serzinc$zinc, conf.int = TRUE, conf.level = 0.95)
tidy(wt)
```

```

# A tibble: 1 x 7
  estimate statistic p.value conf.low conf.high method      alter~1
  <dbl>     <dbl>    <dbl>    <dbl>    <dbl> <chr>      <chr>
1     87.5    106953  2.00e-77     86.0     88.5 Wilcoxon signed rank t~ two.si~
# ... with abbreviated variable name 1: alternative

```

16.28 General Advice

We have described several approaches to estimating a confidence interval for the center of a distribution of quantitative data.

1. The most commonly used approach uses the t distribution to estimate a confidence interval for a population/process mean. This requires some extra assumptions, most particularly that the underlying distribution of the population values is at least approximately Normally distributed. This is identical to the result we get from an intercept-only linear regression model.
2. A more modern and very general approach uses the idea of the bootstrap to estimate a confidence for a population/process parameter, which could be a mean, median or other summary statistic. The bootstrap, and the underlying notion of *resampling* is an important idea that lets us avoid some of the assumptions (in particular Normality) that are required by other methods. Bootstrap confidence intervals involve random sampling, so that the actual values obtained will differ a bit across replications.
3. Finally, the Wilcoxon signed-rank method is one of a number of inferential tools which transform the data to their *ranks* before estimating a confidence interval. This avoids some assumptions, but yields inferences about a less-familiar parameter - the pseudo-median.

Most of the time, the **bootstrap** provides a reasonably adequate confidence interval estimate of the population value of a parameter (mean or median, most commonly) from a distribution when our data consists of a single sample of quantitative information.

17 Ibuprofen in Sepsis

17.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

source("data/Love-boost.R")
library(broom)
library(Epi)
library(knitr)
library(janitor)
library(tidyverse)

theme_set(theme_bw())
```

In addition to the `Love-boost.R` script, we will also use the `favstats` function from the `mosaic` package, and several functions from the `ggridges` and `Hmisc` packages.

17.2 The Trial

Our next study is a randomized controlled trial comparing ibuprofen vs. placebo in patients with sepsis, which uses an *independent samples* design to compare two samples of quantitative data. We will be working with a sample from the Ibuprofen in Sepsis study, which is also studied in Dupont (2002). Quoting the abstract from Bernard et al. (1997):

Ibuprofen has been shown to have effects on sepsis in humans, but because of their small samples (fewer than 30 patients), previous studies have been inadequate to assess effects on mortality. We sought to determine whether ibuprofen can alter rates of organ failure and mortality in patients with the sepsis syndrome, how the drug affects the increased metabolic demand in sepsis (e.g., fever, tachypnea, tachycardia, hypoxemia, and lactic acidosis), and what potential adverse effects the drug has in the sepsis syndrome.

In this study, patients meeting specific criteria (including elevated temperature) for a diagnosis of sepsis were recruited if they fulfilled an additional set of study criteria in the intensive care unit at one of seven participating centers.

The full trial involved 455 patients, of which our sample includes 300. 150 of our patients were randomly assigned to the Ibuprofen group and 150 to the Placebo group¹. I picked the `sepsis` sample we will work with excluding patients with missing values for our outcome of interest, and then selected a random sample of 150 Ibuprofen and 150 Placebo patients from the rest of the group, and converted the temperatures and changes from Fahrenheit to Celsius. The data are gathered in the `sepsis` data file.

```
sepsis <- read_csv("data/sepsis.csv")
```

For the moment, we focus on two variables:

- `treat`, which specifies the treatment group (intravenous Ibuprofen or intravenous Placebo), which was assigned via randomization to each patient, and
- `temp_drop`, the outcome of interest, measured as the change from baseline to 2 hours later in degrees Celsius. Positive values indicate improvement, that is, a *drop* in temperature over the 2 hours following the baseline measurement.

The `sepsis.csv` file also contains each subject's

- `id`, which is just a code
- `race` (three levels: White, AfricanA or Other)
- `apache` = baseline APACHE II score, a severity of disease score ranging from 0 to 71 with higher scores indicating more severe disease and a higher mortality risk
- `temp_0` = baseline temperature, degrees Celsius.

but for the moment, we won't worry about those.

```
sepsis <- sepsis |>  
  mutate(treat = factor(treat),  
         race = factor(race))  
  
summary(sepsis)
```

	<code>id</code>	<code>treat</code>	<code>race</code>	<code>apache</code>
Length:	300	Ibuprofen:150	AfricanA: 80	Min. : 0.0
Class :	character	Placebo :150	Other : 23	1st Qu.:10.0
Mode :	character		White :197	Median :14.0

¹This was a *double-blind* study, where neither the patients nor their care providers know, during the execution of the trial, what intervention group was assigned to each patient.

		Mean :15.4
		3rd Qu.:20.0
		Max. :35.0
temp_0	temp_drop	
Min. :33.10	Min. :-2.7000	
1st Qu.:37.48	1st Qu.:-0.1000	
Median :38.20	Median : 0.3000	
Mean :38.00	Mean : 0.3083	
3rd Qu.:38.70	3rd Qu.: 0.7000	
Max. :41.70	Max. : 3.1000	

17.3 Comparing Two Groups

In making a choice between two alternatives, questions such as the following become paramount.

- Is there a status quo?
- Is there a standard approach?
- What are the costs of incorrect decisions?
- Are such costs balanced?

The process of comparing the means/medians/proportions/rates of the populations represented by two independently obtained samples can be challenging, and such an approach is not always the best choice. Often, specially designed experiments can be more informative at lower cost (i.e. smaller sample size). As one might expect, using these more sophisticated procedures introduces trade-offs, but the costs are typically small relative to the gain in information.

When faced with such a comparison of two alternatives, a test based on **paired** data is often much better than a test based on two distinct, independent samples. Why? If we have done our experiment properly, the pairing lets us eliminate background variation that otherwise hides meaningful differences.

17.3.1 Model-Based Comparisons and ANOVA/Regression

Comparisons based on independent samples of quantitative variables are also frequently accomplished through other equivalent methods, including the analysis of variance approach and dummy variable regression, both of which produce identical confidence intervals to the pooled variance t test for the same comparison.

We will also discuss some of the main ideas in developing, designing and analyzing statistical experiments, specifically in terms of making comparisons. The ideas we will present in this

section allow for the comparison of more than two populations in terms of their population means. The statistical techniques employed analyze the sample variance in order to test and estimate the population means and for this reason the method is called the analysis of variance (ANOVA), and we will discuss this approach alone, and within the context of a linear regression model using dummy or indicator variables.

17.4 Key Questions for Comparing with Independent Samples

17.4.1 What is the population under study?

- All patients in the intensive care unit with sepsis who meet the inclusion and exclusion criteria of the study, at the entire population of health centers like the ones included in the trial.

17.4.2 What is the sample? Is it representative of the population?

- The sample consists of 300 patients. It is a convenient sample from the population under study.
- This is a randomized clinical trial. 150 of the patients were assigned to Ibuprofen, and the rest to Placebo. It is this treatment assignment that is randomized, not the selection of the sample as a whole.
- In expectation, randomization of individuals to treatments, as in this study, should be expected to eliminate treatment selection bias.

17.4.3 Who are the subjects / individuals within the sample?

- 150 patients who received Ibuprofen and a completely different set of 150 patients who received Placebo.
- There is no match or link between the patients. They are best thought of as independent samples.

17.4.4 What data are available on each individual?

- The key variables are the treatment indicator (Ibuprofen or Placebo) and the outcome (drop in temperature in the 2 hours following administration of the randomly assigned treatment.)

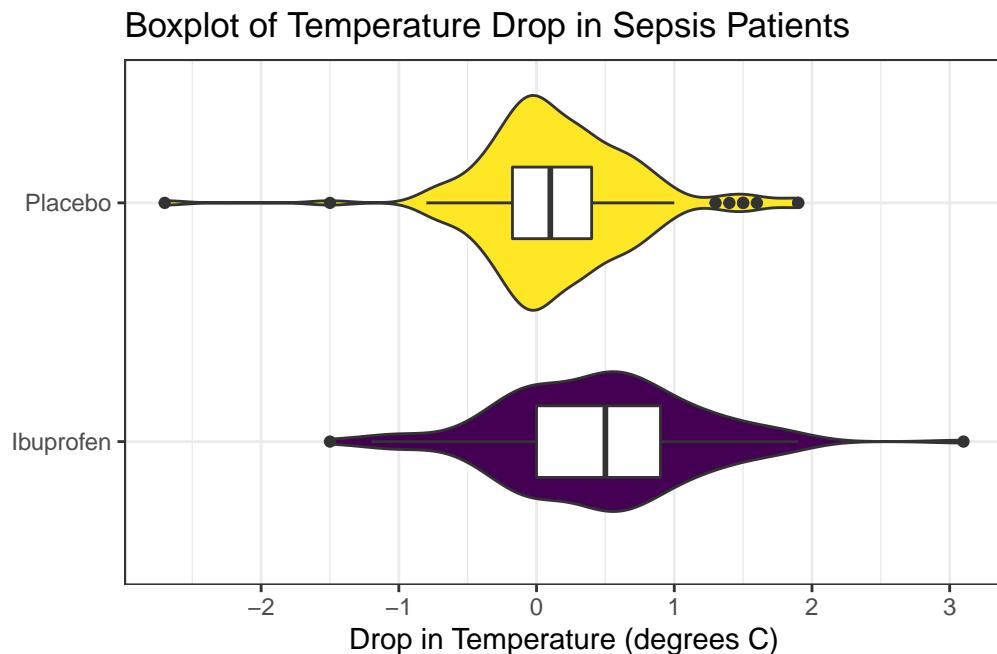
17.4.5 RCT Caveats

The placebo-controlled, double-blind randomized clinical trial, especially if pre-registered, is often considered the best feasible study for assessing the effectiveness of a treatment. While that's not always true, it is a very solid design. The primary caveat is that the patients who are included in such trials are rarely excellent representations of the population of potentially affected patients as a whole.

17.5 Exploratory Data Analysis

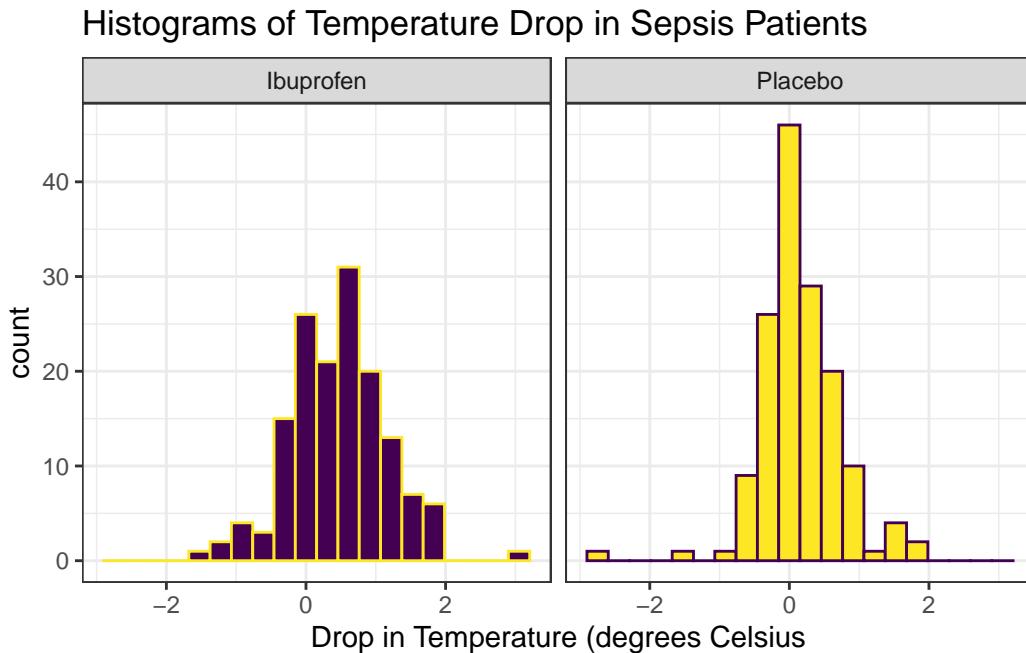
Consider the following boxplot with violin of the `temp_drop` data within each `treat` group.

```
ggplot(sepsis, aes(x = treat, y = temp_drop, fill = treat)) +  
  geom_violin() +  
  geom_boxplot(width = 0.3, fill = "white") +  
  scale_fill_viridis_d() +  
  guides(fill = "none") +  
  labs(title = "Boxplot of Temperature Drop in Sepsis Patients",  
       x = "", y = "Drop in Temperature (degrees C)") +  
  coord_flip()
```



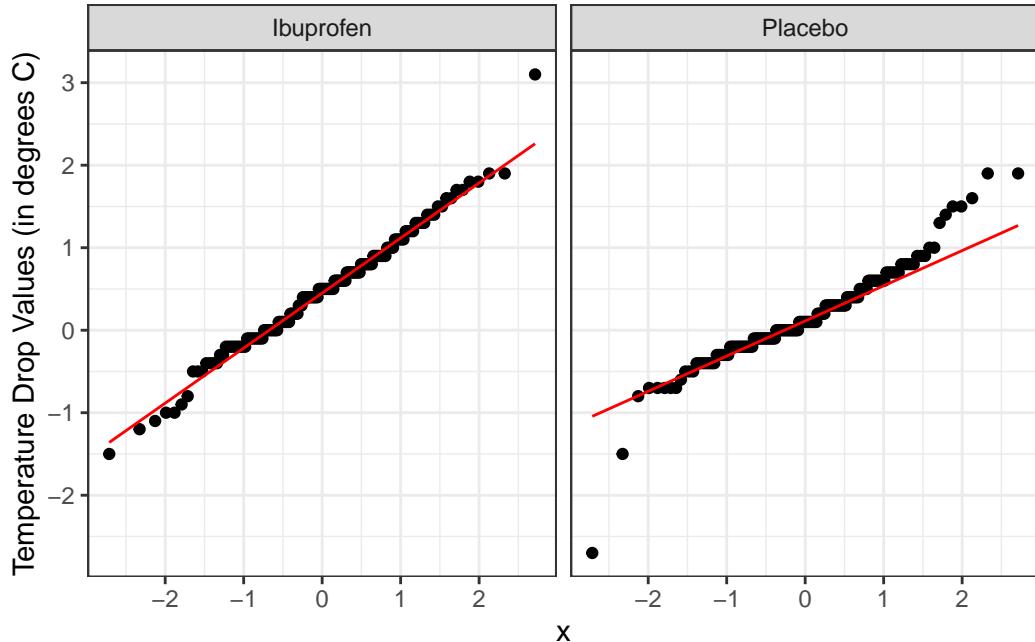
Next, we'll consider faceted histograms of the data.

```
ggplot(sepsis, aes(x = temp_drop, fill = treat, color = treat)) +  
  geom_histogram(bins = 20) +  
  scale_fill_viridis_d() +  
  scale_color_viridis_d(direction = -1) +  
  guides(fill = "none", color = "none") +  
  labs(title = "Histograms of Temperature Drop in Sepsis Patients",  
       x = "Drop in Temperature (degrees Celsius)") +  
  facet_wrap(~ treat)
```



Here's a pair of Normal Q-Q plots. It's not hard to use a Normal model to approximate the Ibuprofen data, but such a model is probably not a good choice for the Placebo results.

```
ggplot(sepsis, aes(sample = temp_drop)) +  
  geom_qq() + geom_qq_line(col = "red") +  
  facet_wrap(~ treat) +  
  labs(y = "Temperature Drop Values (in degrees C)")
```

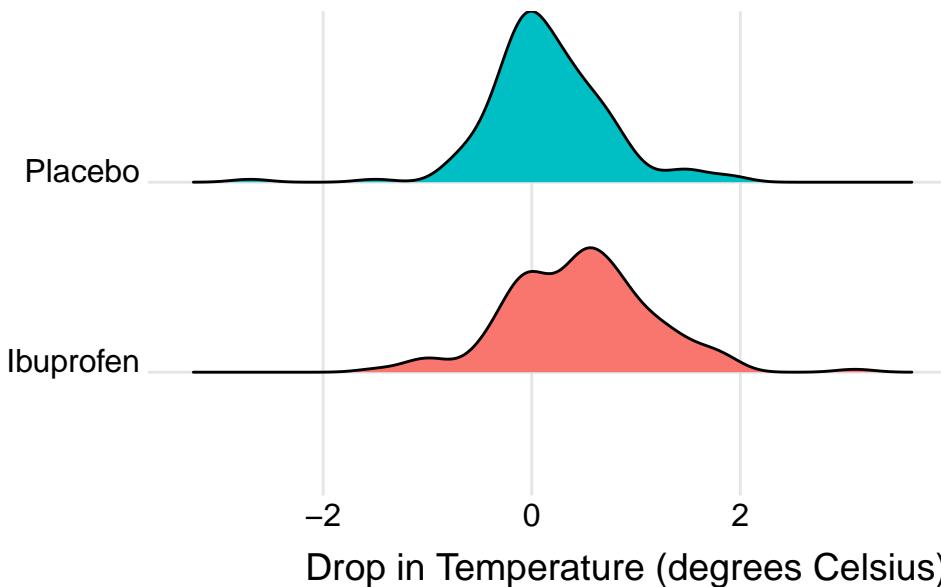


We'll could perhaps also look at a ridgeline plot.

```
ggplot(sepsis, aes(x = temp_drop, y = treat, fill = treat)) +
  ggridges::geom_density_ridges(scale = 0.9) +
  guides(fill = "none") +
  labs(title = "Temperature Drop in Sepsis Patients",
       x = "Drop in Temperature (degrees Celsius)", y = "") +
  ggridges::theme_ridges()
```

Picking joint bandwidth of 0.182

Temperature Drop in Sepsis Patients



The center of the ibuprofen distribution is shifted a bit towards the more positive (greater improvement) direction, it seems, than is the distribution for the placebo patients. This conclusion matches what we see in some key numerical summaries, within the treatment groups.

```
mosaic::favstats(temp_drop ~ treat, data = sepsis)
```

```
Registered S3 method overwritten by 'mosaic':  
  method           from  
fortify.SpatialPolygonsDataFrame ggplot2
```

	treat	min	Q1	median	Q3	max	mean	sd	n	missing
1	Ibuprofen	-1.5	0.000	0.5	0.9	3.1	0.4640000	0.6877919	150	0
2	Placebo	-2.7	-0.175	0.1	0.4	1.9	0.1526667	0.5709637	150	0

17.6 Estimating the Difference in Population Means

Next, we will build a point estimate and 90% confidence interval for the difference between the mean `temp_drop` if treated with Ibuprofen and the mean `temp_drop` if treated with Placebo. We'll use a regression model with a single predictor (the `treat` group) to do this.

```

model_sep <- lm(temp_drop ~ treat == "Ibuprofen", data = sepsis)

tidy(model_sep, conf.int = TRUE, conf.level = 0.90) |>
  kable(digits = 3)

```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.153	0.052	2.958	0.003	0.068	0.238
treat == "Ibuprofen"	0.311	0.073	4.266	0.000	0.191	0.432
TRUE						

The point estimate for the “Ibuprofen - Placebo” difference in population means is 0.311 degrees C, and the 90% confidence interval is (0.191, 0.432) degrees C.

We could also have run the model like this:

```

model_sep2 <- lm(temp_drop ~ treat, data = sepsis)

tidy(model_sep2, conf.int = TRUE, conf.level = 0.90) |>
  kable(digits = 3)

```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.464	0.052	8.991	0	0.379	0.549
treatPlacebo	-0.311	0.073	-4.266	0	-0.432	-0.191

and would therefore conclude that the *Placebo - Ibuprofen* difference was estimated as -0.311, with 90% confidence interval (-0.432, -0.191), which is of course equivalent to our previous estimate.

Fundamentally, this regression model approach is identical to a **two-sample t test, assuming equal population variances**, also called a **pooled t test**. This is just one possible way for us to estimate the difference between population means, as it turns out.

17.7 t-based CI for population mean1 - mean2 difference

17.7.1 The Pooled t procedure

The most commonly used t-procedure for building a confidence interval assumes not only that each of the two populations being compared follows a Normal distribution, but also that they

have the same population variance. This is the pooled t-test, and it is what people usually mean when they describe a two-sample t test.

```
t.test(temp_drop ~ treat,
       data = sepsis,
       conf.level = 0.90,
       alt = "two.sided",
       var.equal = TRUE)
```

Two Sample t-test

```
data: temp_drop by treat
t = 4.2656, df = 298, p-value = 2.68e-05
alternative hypothesis: true difference in means between group Ibuprofen and group Placebo is
90 percent confidence interval:
0.1909066 0.4317600
sample estimates:
mean in group Ibuprofen   mean in group Placebo
0.4640000                 0.1526667
```

Or, we can use `tidy` on this object:

```
tt1 <- t.test(temp_drop ~ treat,
               data = sepsis,
               conf.level = 0.90,
               alt = "two.sided",
               var.equal = TRUE)

tidy(tt1)

# A tibble: 1 x 10
  estim~1 estim~2 estim~3 stati~4 p.value param~5 conf.~6 conf.~7 method alter~8
  <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <chr>  <chr>
1 0.311    0.464    0.153    4.27  2.68e-5     298    0.191  0.432 Two S~ two.si~
# ... with abbreviated variable names 1: estimate, 2: estimate1, 3: estimate2,
# 4: statistic, 5: parameter, 6: conf.low, 7: conf.high, 8: alternative
```

17.7.2 Using linear regression to obtain a pooled t confidence interval

As we've seen, and will demonstrate again below, a linear regression model, using the same outcome and predictor (group) as the pooled t procedure, produces the same confidence inter-

val, again, under the assumption that the two populations we are comparing follow a Normal distribution with the same (population) variance.

```
model1 <- lm(temp_drop ~ treat, data = sepsis)

tidy(model1, conf.int = TRUE, conf.level = 0.90)

# A tibble: 2 x 7
  term      estimate std.error statistic p.value conf.low conf.high
  <chr>     <dbl>     <dbl>     <dbl>    <dbl>    <dbl>    <dbl>
1 (Intercept) 0.464     0.0516    8.99  2.91e-17   0.379    0.549
2 treatPlacebo -0.311    0.0730   -4.27  2.68e- 5  -0.432   -0.191
```

We see that our point estimate from the linear regression model is that the difference in `temp_drop` is -0.3113333, where Ibuprofen subjects have higher `temp_drop` values than do Placebo subjects, and that the 90% confidence interval for this difference ranges from -0.43176 to -0.1909066.

We can obtain a t-based confidence interval for each of the parameter estimates in a linear model directly using `tidy` from the `broom` package. Linear models usually summarize only the estimate and standard error. Remember that a reasonable approximation in large samples to a 95% confidence interval for a regression estimate (slope or intercept) can be obtained from estimate plus or minus two times the standard error.

```
tidy(model1, conf.int = TRUE, conf.level = 0.95) |> kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.464	0.052	8.991	0	0.362	0.566
treatPlacebo	-0.311	0.073	-4.266	0	-0.455	-0.168

So, in the case of the `treatPlacebo` estimate, we can obtain an approximate 95% confidence interval with (-0.457, -0.165). Compare this to the 95% confidence interval available from the model directly, shown in the tidied output above, or with the `confint` command below, and you'll see only a small difference.

Note that we can also use `summary` and `confint` to build our estimates.

```
summary(model1)
```

```

Call:
lm(formula = temp_drop ~ treat, data = sepsis)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.85267 -0.36400 -0.05267  0.34733  2.63600 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.46400   0.05161   8.991 < 2e-16 ***
treatPlacebo -0.31133   0.07299  -4.266 2.68e-05 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6321 on 298 degrees of freedom
Multiple R-squared:  0.05755, Adjusted R-squared:  0.05438 
F-statistic: 18.2 on 1 and 298 DF, p-value: 2.68e-05

```

```
confint(model1, level = 0.95)
```

	2.5 %	97.5 %
(Intercept)	0.3624351	0.5655649
treatPlacebo	-0.4549679	-0.1676988

17.7.3 The Welch t procedure

The default confidence interval based on the t test for independent samples in R uses something called the Welch test, in which the two populations being compared are not assumed to have the same variance. Each population is assumed to follow a Normal distribution.

```
t.test(temp_drop ~ treat, data = sepsis, conf.level = 0.90, alt = "two.sided")
```

```

Welch Two Sample t-test

data: temp_drop by treat
t = 4.2656, df = 288.24, p-value = 2.706e-05
alternative hypothesis: true difference in means between group Ibuprofen and group Placebo is
90 percent confidence interval:
```

```

0.1908939 0.4317728
sample estimates:
mean in group Ibuprofen   mean in group Placebo
0.4640000                  0.1526667

```

Tidying works in this situation, too.

```

tt0 <- t.test(temp_drop ~ treat,
               data = sepsis, conf.level = 0.90, alt = "two.sided")

tidy(tt0)

# A tibble: 1 x 10
estim~1 estim~2 estim~3 stati~4 p.value param~5 conf.~6 conf.~7 method alter~8
<dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <chr>  <chr>
1 0.311    0.464    0.153    4.27 2.71e-5    288.   0.191   0.432 Welch~ two.si~
# ... with abbreviated variable names 1: estimate, 2: estimate1, 3: estimate2,
# 4: statistic, 5: parameter, 6: conf.low, 7: conf.high, 8: alternative

```

When there is a *balanced design*, that is, when the same number of observations appear in each of the two samples, then the Welch t test and the Pooled t test produce the same confidence interval. Differences appear if the sample sizes in the two groups being compared are different.

17.8 Wilcoxon-Mann-Whitney “Rank Sum” CI

As in the one-sample case, a rank-based alternative attributed to Wilcoxon (and sometimes to Mann and Whitney) provides a two-sample comparison of the pseudomedians in the two **treat** groups in terms of **temp_drop**. This is called a **rank sum** test, rather than the Wilcoxon **signed rank** test that is used for inference about a single sample. Here’s the resulting 90% confidence interval for the difference in pseudomedians.

```

wt <- wilcox.test(temp_drop ~ treat, data = sepsis,
                   conf.int = TRUE, conf.level = 0.90,
                   alt = "two.sided")

wt

```

```

Wilcoxon rank sum test with continuity correction

data: temp_drop by treat
W = 14614, p-value = 7.281e-06
alternative hypothesis: true location shift is not equal to 0
90 percent confidence interval:
0.1999699 0.4000330
sample estimates:
difference in location
0.3000368

```

```
tidy(wt)
```

```

# A tibble: 1 x 7
  estimate statistic   p.value conf.low conf.high method      alter~1
    <dbl>     <dbl>     <dbl>     <dbl>     <dbl> <chr>      <chr>
1     0.300    14614. 0.00000728     0.200     0.400 Wilcoxon rank sum te~ two.si~
# ... with abbreviated variable name 1: alternative

```

17.9 Bootstrapping: A More Robust Approach

Within a script called `Love-boost.R`, I have provided the following R code to create a function called `bootdif`.

```

bootdif <-
  function(y, g, conf.level=0.95, B.reps = 2000) {
    lowq = (1 - conf.level)/2
    g <- as.factor(g)
    a <- attr(Hmisc::smean.cl.boot(y[g==levels(g)[1]], B=B.reps, reps=TRUE), 'reps')
    b <- attr(Hmisc::smean.cl.boot(y[g==levels(g)[2]], B=B.reps, reps=TRUE), 'reps')
    meandif <- diff(tapply(y, g, mean, na.rm=TRUE))
    a.b <- quantile(b-a, c(lowq, 1-lowq))
    res <- c(meandif, a.b)
    names(res) <- c('Mean Difference', lowq, 1-lowq)
    res
  }

```

Running this code will place a new function called `bootdif` in your environment, which will help us calculate an appropriate confidence interval using a bootstrap procedure. The `bootdif`

function contained in the `Love-boost.R` script is a slightly edited version of the function at <http://biostat.mc.vanderbilt.edu/wiki/Main/BootstrapMeansSoftware>.

17.9.1 Bootstrap CI for the Sepsis study

Note that this approach uses a comma to separate the outcome variable (here, `temp_drop`) from the variable identifying the exposure groups (here, `treat`).

```
set.seed(431212)

bootdif(sepsis$temp_drop, sepsis$treat, conf.level = 0.90)
```

Mean Difference	0.05	0.95
-0.3113333	-0.4313667	-0.1833000

This approach calculates a 90% confidence interval for the difference in means between the two treatment groups. Note that the sign is in the opposite direction from what we've seen in our previous work. We can tell from the mean difference (and the summarized means from the data in each group) that this approach is finding a confidence interval using a bootstrap procedure for the Placebo - Ibuprofen difference, specifically (-0.431, -0.183).

```
mosaic::favstats(temp_drop ~ treat, data = sepsis)
```

	treat	min	Q1	median	Q3	max	mean	sd	n	missing
1	Ibuprofen	-1.5	0.000	0.5	0.9	3.1	0.4640000	0.6877919	150	0
2	Placebo	-2.7	-0.175	0.1	0.4	1.9	0.1526667	0.5709637	150	0

To find a confidence interval using this bootstrap approach for the Ibuprofen - Placebo difference, we just need to switch the signs, and conclude that the 90% bootstrap confidence interval for that difference would be (0.183, 0.431).

17.10 Summary: Specifying A Two-Sample Study Design

These questions will help specify the details of the study design involved in any comparison of two populations on a quantitative outcome, perhaps with means.

1. What is the outcome under study?
2. What are the (in this case, two) treatment/exposure groups?
3. Were the data collected using matched / paired samples or independent samples?

4. Are the data a random sample from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the sample to the population(s)?
5. What is the significance level (or, the confidence level) we require here?
6. Are we doing one-sided or two-sided testing/confidence interval generation?
7. If we have paired samples, did pairing help reduce nuisance variation?
8. If we have paired samples, what does the distribution of sample paired differences tell us about which inferential procedure to use?
9. If we have independent samples, what does the distribution of each individual sample tell us about which inferential procedure to use?

17.11 Results for the sepsis study

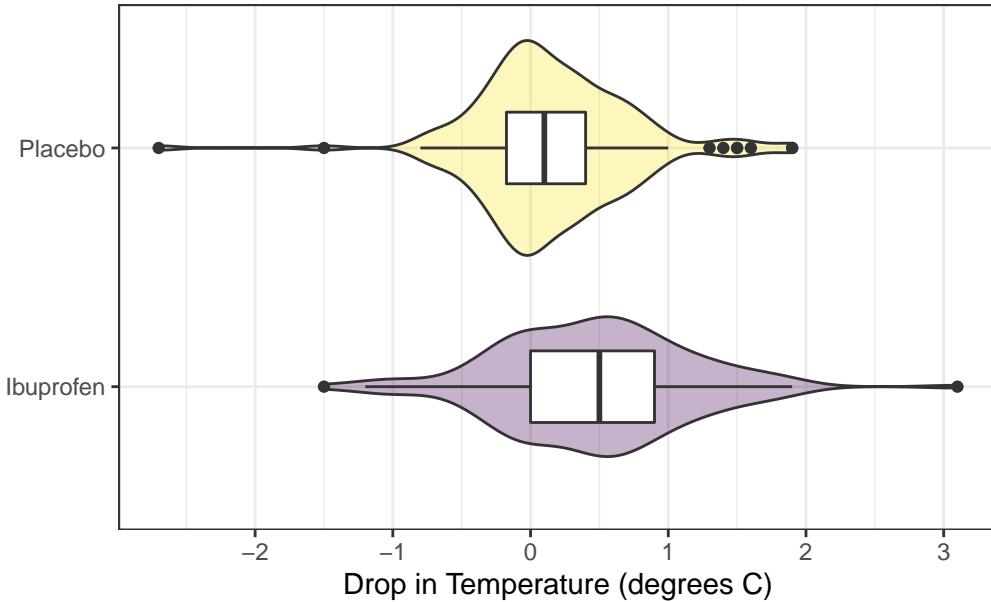
1. The outcome is `temp_drop`, the change in body temperature (in °C) from baseline to 2 hours later, so that positive numbers indicate drops in temperature (a good outcome.)
2. The groups are **Ibuprofen** and **Placebo** as contained in the `treat` variable in the `sepsis` tibble.
3. The data were collected using independent samples. The Ibuprofen subjects are not matched or linked to individual Placebo subjects - they are separate groups.
4. The subjects of the study aren't drawn from a random sample of the population of interest, but they are randomly assigned to their respective treatments (Ibuprofen and Placebo) which will provide the reasoned basis for our inferences.
5. We'll use a 10% significance level (or 90% confidence level) in this setting, as we did in our previous work on these data.
6. We'll use a two-sided testing and confidence interval approach.

Questions 7 and 8 don't apply, because these are independent samples of data, rather than paired samples.

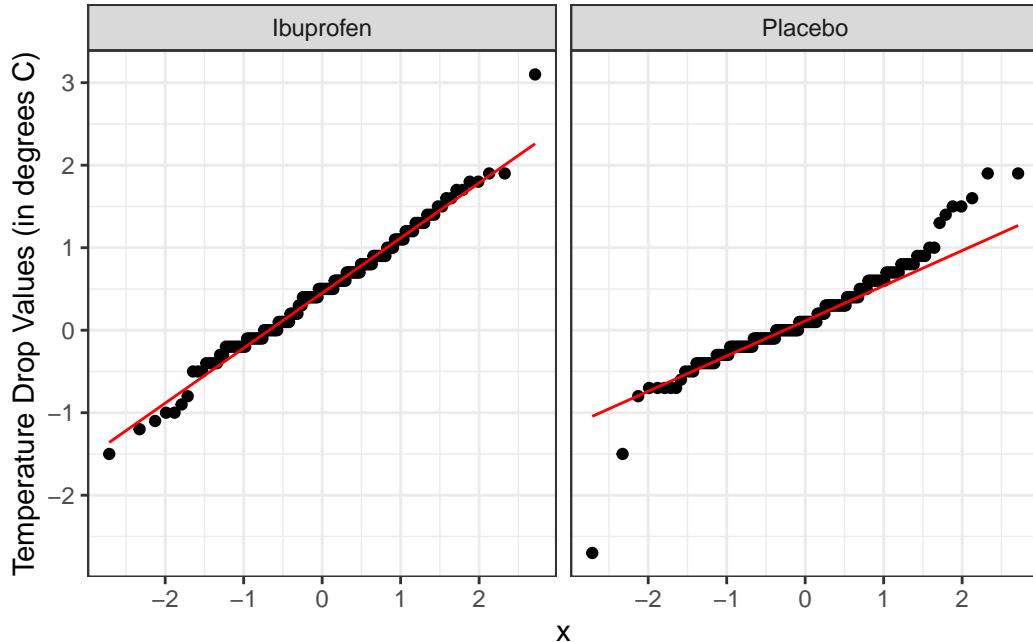
To address question 9, we'll need to look at the data in each sample, as we did previously to allow us to assess the Normality of the distributions of (separately) the `temp_drop` results in the Ibuprofen and Placebo groups. We'll repeat those below.

```
ggplot(sepsis, aes(x = treat, y = temp_drop, fill = treat)) +
  geom_violin() +
  geom_boxplot(width = 0.3, fill = "white") +
  scale_fill_viridis_d(alpha = 0.3) +
  guides(fill = "none") +
  labs(title = "Boxplot of Temperature Drop in Sepsis Patients",
       x = "", y = "Drop in Temperature (degrees C)") +
  coord_flip()
```

Boxplot of Temperature Drop in Sepsis Patients



```
ggplot(sepsis, aes(sample = temp_drop)) +  
  geom_qq() + geom_qq_line(col = "red") +  
  facet_wrap(~ treat) +  
  labs(y = "Temperature Drop Values (in degrees C)")
```



From these plots we conclude that the data in the Ibuprofen sample follow a reasonably Normal distribution, but this isn't quite as true for the Placebo sample. It's hard to know whether the apparent Placebo group outliers will affect whether the Normal distribution assumption is reasonable, so we can see if the confidence intervals change much when we *don't* assume Normality (for instance, comparing the bootstrap to the t-based approaches), as a way of understanding whether a Normal model has a large impact on our conclusions.

17.11.1 Sepsis Estimation Results

Here's a set of confidence interval estimates (we'll use 90% confidence here) using the methods discussed in this Chapter.

```
mosaic::favstats(temp_drop ~ treat, data = sepsis)

  treat   min     Q1 median   Q3 max      mean        sd    n missing
1 Ibuprofen -1.5  0.000    0.5 0.9 3.1 0.4640000 0.6877919 150      0
2 Placebo   -2.7 -0.175    0.1 0.4 1.9 0.1526667 0.5709637 150      0

s_pooled_t_test <- t.test(temp_drop ~ treat, data = sepsis, conf.level = 0.90,
                           alt = "two.sided", var.equal = TRUE)
```

```

tidy(s_pooled_t_test) |>
  select(conf.low, conf.high)

# A tibble: 1 x 2
  conf.low conf.high
  <dbl>     <dbl>
1     0.191     0.432

s_welch_t_test <- t.test(temp_drop ~ treat, data = sepsis, conf.level = 0.90,
                         alt = "two.sided", var.equal = FALSE)

tidy(s_welch_t_test) |>
  select(estimate, conf.low, conf.high)

# A tibble: 1 x 3
  estimate conf.low conf.high
  <dbl>     <dbl>     <dbl>
1     0.311     0.191     0.432

s_wilcoxon_test <- wilcox.test(temp_drop ~ treat, data = sepsis,
                                 conf.int = TRUE, conf.level = 0.90,
                                 alt = "two.sided")

tidy(s_wilcoxon_test) |>
  select(estimate, conf.low, conf.high)

# A tibble: 1 x 3
  estimate conf.low conf.high
  <dbl>     <dbl>     <dbl>
1     0.300     0.200     0.400

set.seed(431212)
s_bootstrap <- bootdif(sepsis$temp_drop, sepsis$treat, conf.level = 0.90)

s_bootstrap
```

Mean Difference	0.05	0.95
-0.3113333	-0.4313667	-0.1833000

Procedure	Compares...	Point Estimate	90% CI
Pooled t	Means	0.311	(0.191, 0.432)
Welch t	Means	0.311	(0.191, 0.432)
Bootstrap	Means	0.311	(0.183, 0.431)
Wilcoxon rank sum	Pseudo-Medians	0.3	(0.2, 0.4)

What conclusions can we draw in this setting?

17.12 Categorizing the Outcome and Comparing Rates

Suppose we were interested in comparing the percentage of patients in each arm of the trial (Ibuprofen vs. Placebo) that showed an improvement in their temperature (`temp_drop > 0`). To build the cross-tabulation of interest, we could create a new variable, called `dropped` which indicates whether the subject's temperature dropped, and then use `tabyl`.

```
sepsis <- sepsis |>
  mutate(dropped = ifelse(temp_drop > 0, "Drop", "No Drop"))

sepsis |> tabyl(treat, dropped)
```

	treat	Drop	No Drop
Ibuprofen	107	43	
Placebo	80	70	

Our primary interest is in comparing the percentage of Ibuprofen patients whose temperature dropped to the percentage of Placebo patients whose temperature dropped.

```
sepsis |> tabyl(treat, dropped) |>
  adorn_totals() |>
  adorn_percentages(denom = "row") |>
  adorn_pct_formatting(digits = 1) |>
  adorn_ns(position = "front")

  treat      Drop      No Drop
Ibuprofen 107 (71.3%) 43 (28.7%)
  Placebo   80 (53.3%) 70 (46.7%)
  Total    187 (62.3%) 113 (37.7%)
```

17.13 Estimating the Difference in Proportions

In our sample, 71.3% of the Ibuprofen subjects, and 53.3% of the Placebo subjects, experienced a drop in temperature. So our *point estimate* of the difference in percentages would be 18.0 percentage points, but we will usually set this instead in terms of proportions, so that the difference is 0.180.

Now, we'll find a confidence interval for that difference, which we can do in several ways, including the `twoby2` function in the Epi package.

```
table(sepsis$treat, sepsis$dropped) |> twoby2(alpha = 0.10)
```

2 by 2 table analysis:

Outcome : Drop

Comparing : Ibuprofen vs. Placebo

	Drop	No Drop	P(Drop)	90% conf. interval
Ibuprofen	107	43	0.7133	0.6490 0.7701
Placebo	80	70	0.5333	0.4661 0.5993

90% conf. interval

Relative Risk: 1.3375 1.1492 1.5567

Sample Odds Ratio: 2.1773 1.4583 3.2509

Conditional MLE Odds Ratio: 2.1716 1.4177 3.3437

Probability difference: 0.1800 0.0881 0.2677

Exact P-value: 0.0019

Asymptotic P-value: 0.0014

While there is a lot of additional output here, we'll look for now just at the Probability difference row, where we see the point estimate (0.180) and the 90% confidence interval estimate for the difference in proportions (0.088, 0.268) comparing Ibuprofen vs. Placebo for the outcome of Dropping in Temperature.

More on estimation of the difference in population proportions will be found later.

18 Comparing Means with Paired Samples

Here, we'll consider the problem of estimating a confidence interval to describe the difference in population means (or medians) based on a comparison of two samples of quantitative data, gathered using a matched pairs design.

Specifically, we'll use as our example the Lead in the Blood of Children study, described below.

18.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

source("data/Love-boost.R")
library(knitr)
library(broom)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

In addition to the `Love-boost.R` script, we will also use the `favstats` function from the `mosaic` package, and several functions from the `ggridges` and `Hmisc` packages.

18.2 Lead in the Blood of Children

One of the best ways to eliminate a source of variation and the errors of interpretation associated with it is through the use of matched pairs. Each subject in one group is matched as closely as possible by a subject in the other group. If a 45-year-old African-American male with hypertension is given a [treatment designed to lower their blood pressure], then we give a second, similarly built 45-year old African-American male with hypertension a placebo.

- Good (2005), section 5.2.4

18.3 The Lead in the Blood of Children Study

Morton et al. (1982) studied the absorption of lead into the blood of children. This was a matched-sample study, where the exposed group of interest contained 33 children of parents who worked in a battery manufacturing factory (where lead was used) in the state of Oklahoma. Specifically, each child with a lead-exposed parent was matched to another child of the same age, exposure to traffic, and living in the same neighborhood whose parents did not work in lead-related industries. So the complete study had 66 children, arranged in 33 matched pairs. The outcome of interest, gathered from a sample of whole blood from each of the children, was lead content, measured in mg/dl.

One motivation for doing this study is captured in the Abstract from Morton et al. (1982).

It has been repeatedly reported that children of employees in a lead-related industry are at increased risk of lead absorption because of the high levels of lead found in the household dust of these workers.

The data are available in several places, including Table 5 of Pruzek and Helmreich (2009), in the `BloodLead` data set within the `PairedData` package in R, but we also make them available in the `bloodlead.csv` file. A table of the first few pairs of observations (blood lead levels for one child exposed to lead and the matched control) is shown below.

```
bloodlead <- read_csv("data/bloodlead.csv")  
  
Rows: 33 Columns: 3  
-- Column specification -----  
Delimiter: ","  
chr (1): pair  
dbl (2): exposed, control  
  
i Use `spec()` to retrieve the full column specification for this data.  
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
bloodlead
```

```
# A tibble: 33 x 3  
  pair   exposed control  
  <chr>    <dbl>    <dbl>  
1 P01      38      16  
2 P02      23      18  
3 P03      41      18
```

```

4 P04      18      24
5 P05      37      19
6 P06      36      11
7 P07      23      10
8 P08      62      15
9 P09      31      16
10 P10     34      18
# ... with 23 more rows
# i Use `print(n = ...)` to see more rows

```

- In each pair, one child was exposed (to having a parent working in the factory) and the other was not.
- Otherwise, though, each child was very similar to its matched partner.
- The data under **exposed** and **control** are the blood lead content, in mg/dl.

Our primary goal will be to estimate the difference in lead content between the exposed and control children, and then use that sample estimate to make inferences about the difference in lead content between the population of all children like those in the exposed group and the population of all children like those in the control group.

18.3.1 Our Key Questions for a Paired Samples Comparison

1. What is the **population** under study?
 - All pairs of children living in Oklahoma near the factory in question, in which one had a parent working in a factory that exposed them to lead, and the other did not.
2. What is the **sample**? Is it representative of the population?
 - The sample consists of 33 pairs of one exposed and one control child.
 - This is a case-control study, where the children were carefully enrolled to meet the design criteria. Absent any other information, we're likely to assume that there is no serious bias associated with these pairs, and that assuming they represent the population effectively (and perhaps the broader population of kids whose parents work in lead-based industries more generally) may well be at least as reasonable as assuming they don't.
3. Who are the subjects / **individuals** within the sample?
 - Each of our 33 pairs of children includes one exposed child and one unexposed (control) child.
4. What **data** are available on each individual?
 - The blood lead content, as measured in mg/dl of whole blood.

18.3.2 Lead Study Caveats

Note that the children were not randomly selected from general populations of kids whose parents did and did not work in lead-based industries.

- To make inferences to those populations, we must make **strong assumptions** to believe, for instance, that the sample of exposed children is as representative as a random sample of children with similar exposures across the world would be.
- The researchers did have a detailed theory about how the exposed children might be at increased risk of lead absorption, and in fact as part of the study gathered additional information about whether a possible explanation might be related to the quality of hygiene of the parents (all of them were fathers, actually) who worked in the factory.
- This is an observational study, so that the estimation of a causal effect between parental work in a lead-based industry and children's blood lead content can be made, without substantial (and perhaps heroic) assumptions.

18.4 Exploratory Data Analysis for Paired Samples

We'll begin by adjusting the data in two ways.

- We'd like that first variable (`pair`) to be a `factor` rather than a `character` type in R, because we want to be able to summarize it more effectively. So we'll make that change.
- Also, we'd like to calculate the difference in lead content between the exposed and the control children in each pair, and we'll save that within-pair difference in a variable called `lead_diff`. We'll take `lead_diff = exposed - control` so that positive values indicate increased lead in the exposed child.

```
bloodlead_original <- bloodlead

bloodlead <- bloodlead_original |>
  mutate(pair = factor(pair),
         lead_diff = exposed - control)

bloodlead

# A tibble: 33 x 4
  pair   exposed control lead_diff
  <fct>   <dbl>    <dbl>     <dbl>
1 P01      38       16      22
2 P02      23       18       5
3 P03      41       18      23
```

```

4 P04      18     24     -6
5 P05      37     19     18
6 P06      36     11     25
7 P07      23     10     13
8 P08      62     15     47
9 P09      31     16     15
10 P10     34     18     16
# ... with 23 more rows
# i Use `print(n = ...)` to see more rows

```

18.4.1 The Paired Differences

To begin, we focus on `lead_diff` for our exploratory work, which is the exposed - control difference in lead content within each of the 33 pairs. So, we'll have 33 observations, as compared to the 462 in the serum zinc data, but most of the same tools are still helpful.

```

p1 <- ggplot(bloodlead, aes(sample = lead_diff)) +
  geom_qq(col = "darkslategray") + geom_qq_line(col = "navy") +
  theme(aspect.ratio = 1) +
  labs(title = "Normal Q-Q plot")

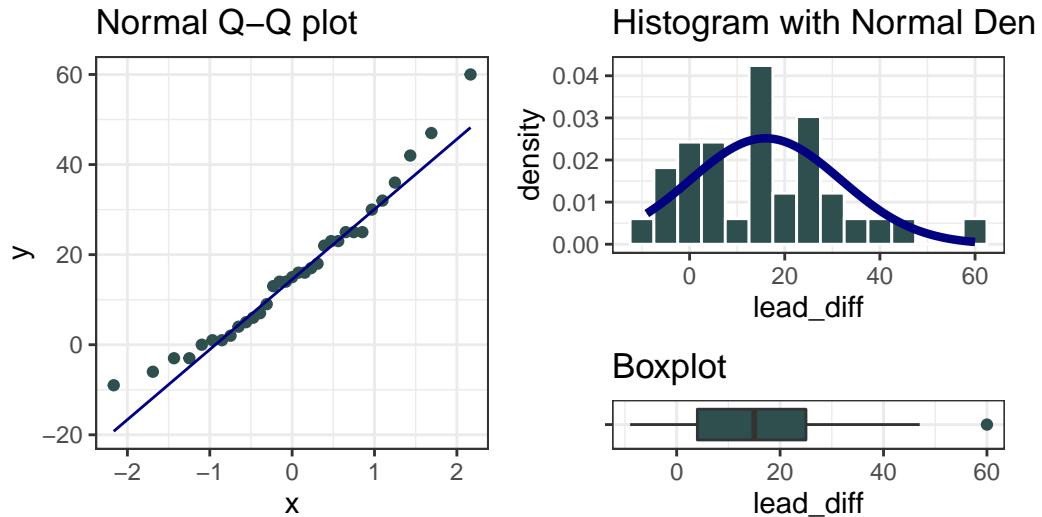
p2 <- ggplot(bloodlead, aes(x = lead_diff)) +
  geom_histogram(aes(y = stat(density)),
                 binwidth = 5, fill = "darkslategray", col = "white") +
  stat_function(fun = dnorm,
                args = list(mean = mean(bloodlead$lead_diff),
                            sd = sd(bloodlead$lead_diff)),
                col = "navy", lwd = 1.5) +
  labs(title = "Histogram with Normal Density")

p3 <- ggplot(bloodlead, aes(x = lead_diff, y = "")) +
  geom_boxplot(fill = "darkslategray", outlier.color = "darkslategray") +
  labs(title = "Boxplot", y = "")

p1 + (p2 / p3 + plot_layout(heights = c(4,1))) +
  plot_annotation(title = "Difference in Blood Lead Content (mg/dl) for 33 Pairs of Children")

```

Difference in Blood Lead Content (mg/dl) for 33 Pairs of Children



Note that in all of this work, I plotted the paired differences. One obvious way to tell if you have paired samples is that you can pair every single subject from one exposure group to a unique subject in the other exposure group. Everyone has to be paired, so the sample sizes will always be the same in the two groups.

Here's a summary of the paired differences.

```
mosaic::favstats(~ lead_diff, data = bloodlead)

min   Q1   median   Q3   max     mean      sd    n missing
-9     4       15    25    60  15.9697  15.86365 33        0

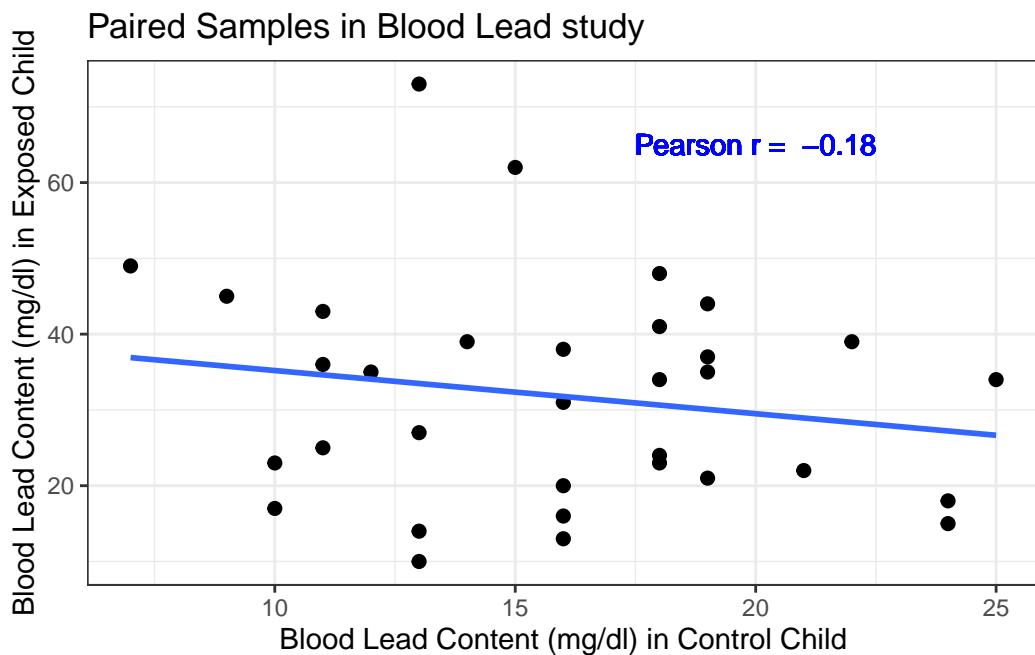
bloodlead |> summarize(skew1 =
  (mean(lead_diff) - median(lead_diff)) /
  sd(lead_diff))

# A tibble: 1 x 1
skew1
<dbl>
1 0.0611
```

18.4.2 Impact of Matching - Scatterplot and Correlation

Here, the data are paired by the study through matching on neighborhood, age and exposure to traffic. Each individual child's outcome value is part of a pair with the outcome value for his/her matching partner. We can see this pairing in several ways, perhaps by drawing a scatterplot of the pairs.

```
ggplot(bloodlead, aes(x = control, y = exposed)) +  
  geom_point(size = 2) +  
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +  
  geom_text(x = 20, y = 65, col = "blue",  
            label =  
            paste("Pearson r = ",  
                  round(cor(bloodlead$control, bloodlead$exposed), 2))) +  
  labs(title = "Paired Samples in Blood Lead study",  
        x = "Blood Lead Content (mg/dl) in Control Child",  
        y = "Blood Lead Content (mg/dl) in Exposed Child")
```



Each point here represents a **pair** of observations, one from a control child, and one from the matched exposed child. If there is a strong linear relationship (usually with a positive slope, thus positive correlation) between the paired outcomes, then the pairing will be more

helpful in terms of improving statistical power of the estimates we build than if there is a weak relationship.

- The stronger the Pearson correlation coefficient, the more helpful pairing will be.
- Here, a straight line model using the control child's blood lead content accounts for about 3.2% of the variation in blood lead content in the exposed child.
- As it turns out, pairing will have only a modest impact here on the inferences we draw in the study. We still will treat the data as paired, despite this.

18.5 Looking at Separate Samples: Using `pivot_longer`

For the purpose of estimating the difference between the exposed and control children, the summaries of the paired differences are what we'll need.

In some settings, however, we might also look at a boxplot, or violin plot, or ridgeline plot that showed the distributions of exposed and control children separately. But we will run into trouble because one variable (blood lead content) is spread across multiple columns (control and exposed.) The solution is to “pivot” the tibble from its current format to build a new, tidy tibble. Because the data aren't *tidied* here, so that we have one row for each subject and one column for each variable, we have to do some work to get them in that form for our usual plotting strategy to work well.

- `pivot_longer()` “lengthens” the data, increasing the number of rows and decreasing the number of columns.
- `pivot_wider()` performs the inverse of that transformation, “widening” the data.

In our original `bloodlead` data, if we drop the `lead_diff` addition we made, we have *wide* data, with each row representing two different subjects.

```
head(bloodlead_original, 3)
```

```
# A tibble: 3 x 3
  pair   exposed control
  <chr>    <dbl>   <dbl>
1 P01      38     16
2 P02      23     18
3 P03      41     18
```

And what we want to accomplish is to have one row for each subject, instead of one row for each pair of subjects. So we want to make the data **longer**.

```

bloodlead_longer <- bloodlead_original |>
  pivot_longer(
    cols = -c(pair),
    names_to = "status",
    values_to = "lead_level")

bloodlead_longer

# A tibble: 66 x 3
  pair   status  lead_level
  <chr> <chr>      <dbl>
1 P01   exposed     38
2 P01   control     16
3 P02   exposed     23
4 P02   control     18
5 P03   exposed     41
6 P03   control     18
7 P04   exposed     18
8 P04   control     24
9 P05   exposed     37
10 P05  control     19
# ... with 56 more rows
# i Use `print(n = ...)` to see more rows

```

For more on this approach (in this case, we're making the data "longer" and its opposite would be making the data "wider"), visit the Tidy data chapter in Wickham and Grolemund (2022) and the `tidyverse` repository on Github at <https://github.com/tidyverse/tidyr>.

And now, we can plot as usual to compare the two samples.

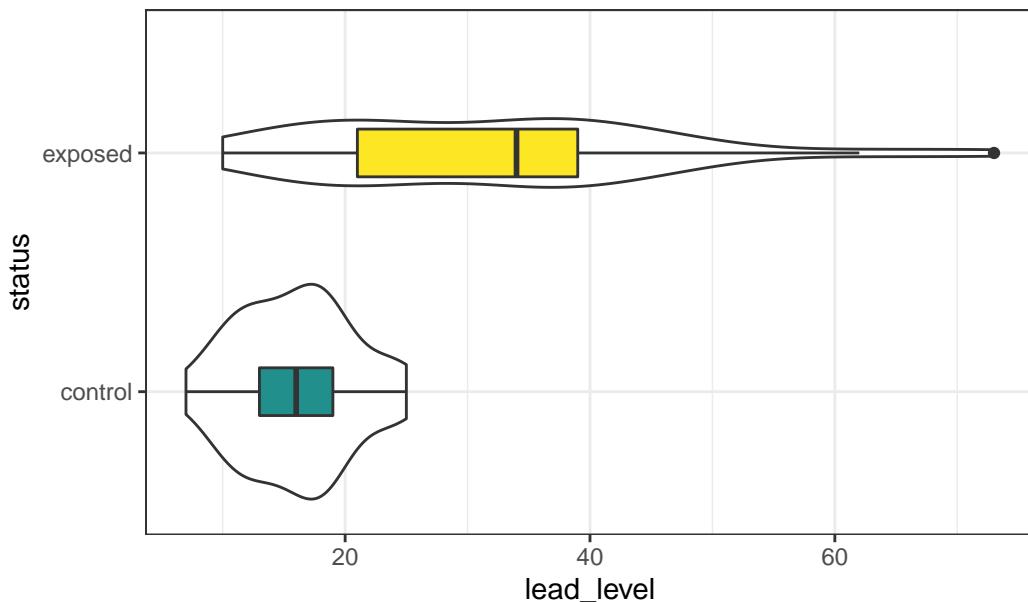
First, we'll look at a boxplot, showing all of the data.

```

ggplot(bloodlead_longer, aes(x = status, y = lead_level)) +
  geom_violin() +
  geom_boxplot(aes(fill = status), width = 0.2) +
  scale_fill_viridis_d(begin = 0.5) +
  guides(fill = "none") +
  coord_flip() +
  labs(title = "Boxplot of Lead Content in Exposed and Control kids")

```

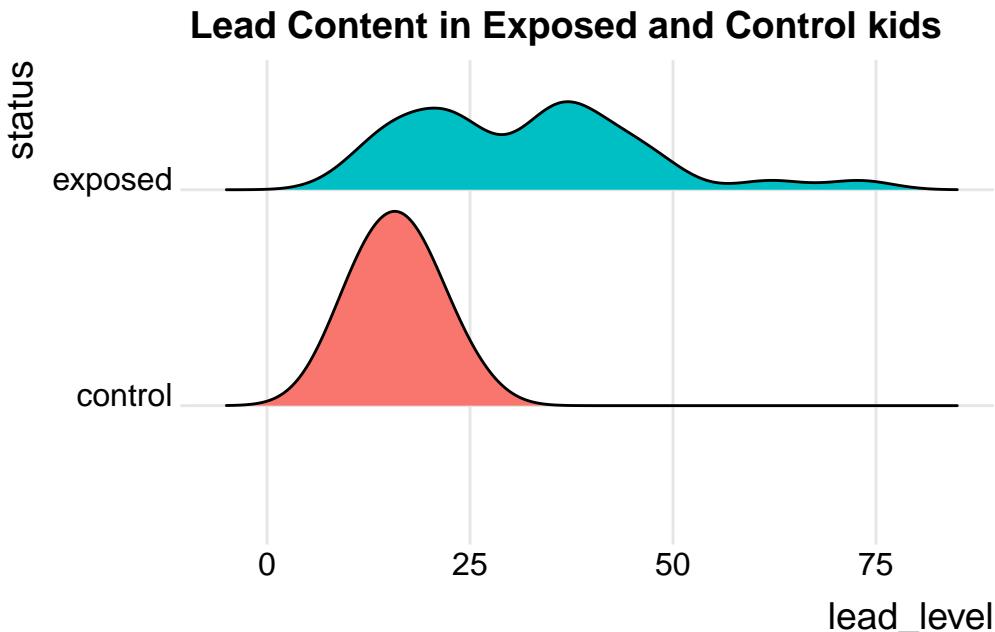
Boxplot of Lead Content in Exposed and Control kids



We'll also look at a ridgeline plot, because Dr. Love likes them, even though they're really more useful when we're comparing more than two samples.

```
ggplot(bloodlead_longer, aes(x = lead_level, y = status, fill = status)) +  
  ggridges::geom_density_ridges(scale = 0.9) +  
  guides(fill = "none") +  
  labs(title = "Lead Content in Exposed and Control kids") +  
  ggridges::theme_ridges()
```

Picking joint bandwidth of 4.01



Both the center and the spread of the distribution are substantially larger in the exposed group than in the controls. Of course, numerical summaries show these patterns, too.

```
mosaic::favstats(lead_level ~ status, data = bloodlead_longer) |>
  kable(digits = 2)
```

status	min	Q1	median	Q3	max	mean	sd	n	missing
control	7	13	16	19	25	15.88	4.54	33	0
exposed	10	21	34	39	73	31.85	14.41	33	0

18.6 Estimating the Difference in Means with Paired Samples

Suppose we want to estimate the difference in the mean blood level across the population of children represented by the sample taken in this study. To do so, we must take advantage of the matched samples design, and complete our estimation on the paired differences, treating them as if they were a single sample of data.

One way to accomplish this is simply to run the usual intercept-only linear regression model on the paired differences.

```

model_lead <- lm(lead_diff ~ 1, data = bloodlead)

tidy(model_lead, conf.int = TRUE, conf.level = 0.90) |>
  kable(digits = 2)

```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	15.97	2.76	5.78	0	11.29	20.65

Our point estimate for the difference (exposed - control) in lead levels is 15.97 mg/dl, and our 90% confidence interval is (11.29, 20.65) mg/dl.

18.6.1 Paired Data in Longer Format?

If we had the data in “longer” format, as in `bloodlead_longer`, with the pairs identified by the `pair` variable, then we could obtain the same confidence interval using:

```

model2_lead <- lm(lead_level ~ status + factor(pair), data = bloodlead_longer)

tidy(model2_lead, conf.int = TRUE, conf.level = 0.90)

```

```

# A tibble: 34 x 7
  term      estimate std.error statistic   p.value conf.low conf.high
  <chr>     <dbl>    <dbl>     <dbl>     <dbl>    <dbl>    <dbl>
1 (Intercept) 19.0     8.05     2.36    0.0244    5.38    32.7
2 statusexposed 16.0     2.76     5.78  0.00000204   11.3    20.6
3 factor(pair)P02 -6.50    11.2    -0.579    0.566   -25.5    12.5
4 factor(pair)P03  2.50     11.2     0.223    0.825   -16.5    21.5
5 factor(pair)P04 -6.00    11.2    -0.535    0.596   -25.0    13.0
6 factor(pair)P05  1.00     11.2     0.0891   0.930   -18.0    20.0
7 factor(pair)P06 -3.50    11.2    -0.312    0.757   -22.5    15.5
8 factor(pair)P07 -10.5    11.2    -0.936    0.356   -29.5     8.50
9 factor(pair)P08  11.5     11.2     1.03     0.313   -7.50    30.5
10 factor(pair)P09 -3.50    11.2    -0.312    0.757  -22.5    15.5
# ... with 24 more rows
# i Use `print(n = ...)` to see more rows

```

and the key elements are found in the `statusexposed` row, which we can focus on nicely (since the output of the `tidy()` function is always a tibble) with:

```

tidy(model2_lead, conf.int = TRUE, conf.level = 0.90) |>
  filter(term == "statusexposed") |>
  kable(digits = 2)

```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
statusexposed	15.97	2.76	5.78	0	11.29	20.65

and again, we have our 90% confidence interval estimate of the population mean difference between exposed and control children.

18.7 Matched Pairs vs. Two Independent Samples

These data were NOT obtained from two independent samples, but rather from matched pairs.

- We only have matched pairs if each individual observation in the “treatment” group is matched to one and only one observation in the “control” group by the way in which the data were gathered. Paired (or matched) data can arise in several ways.
 - The most common is a “pre-post” study where subjects are measured both before and after an exposure happens.
 - In observational studies, we often match up subjects who did and did not receive an exposure so as to account for differences on things like age, sex, race and other covariates. This is what happens in the Lead in the Blood of Children study.
- If the data are from paired samples, we should (and in fact) must form paired differences, with no subject left unpaired.
 - If we cannot line up the data comparing two samples of quantitative data so that the links between the individual “treated” and “control” observations to form matched pairs are evident, then the data are not paired.
 - If the sample sizes were different, we’d know we have independent samples, because matched pairs requires that each subject in the “treated” group be matched to a single, unique member of the “control” group, and thus that we have exactly as many “treated” as “control” subjects.
 - But having as many subjects in one treatment group as the other (which is called a *balanced design*) is only necessary, and not sufficient, for us to conclude that matched pairs are used.

As Bock, Velleman, and De Veaux (2004) suggest,

... if you know the data are paired, you can take advantage of that fact - in fact, you *must* take advantage of it. ... You must decide whether the data are paired from understanding how they were collected and what they mean. ... There is no test to determine whether the data are paired.

18.8 Estimating the Population Mean of the Paired Differences

There are two main approaches used frequently to estimate the population mean of paired differences.

- Estimation using the t distribution (and assuming at least an approximately Normal distribution for the paired differences)
- Estimation using the bootstrap (which doesn't require the Normal assumption)

In addition, we might consider estimating an alternate statistic when the data don't follow a symmetric distribution, like the median, with the bootstrap. In other settings, a rank-based alternative called the Wilcoxon signed rank test is available to estimate a psuedo-median. All of these approaches mirror what we did with a single sample, earlier in these Notes.

18.9 t-based CI for Population Mean of Paired Differences

In R, there are at least five different methods for obtaining the t-based confidence interval for the population difference in means between paired samples. They are all mathematically identical. The key idea is to calculate the paired differences (exposed - control, for example) in each pair, and then treat the result as if it were a single sample and apply the methods developed for that situation earlier in these Notes.

18.9.1 Method 1

We can use the single-sample approach, applied to the variable containing the paired differences. Let's build a **90%** two-sided confidence interval for the population mean of the difference in blood lead content across all possible pairs of an exposed (parent works in a lead-based industry) and a control (parent does not) child.

```
tt1 <- t.test(bloodlead$lead_diff, conf.level = 0.90,  
               alt = "two.sided")
```

```
tt1
```

One Sample t-test

```
data: bloodlead$lead_diff
t = 5.783, df = 32, p-value = 2.036e-06
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
11.29201 20.64738
sample estimates:
mean of x
15.9697
```

```
tidy(tt1) |> kable(digits = 2)
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
15.97	5.78	0	32	11.29	20.65	One Sample t-test	two.sided

The 90% confidence interval is (11.29, 20.65) according to this t-based procedure. An appropriate interpretation of the 90% two-sided confidence interval would be:

- (11.29, 20.65) milligrams per deciliter is a 90% two-sided confidence interval for the population mean difference in blood lead content between exposed and control children.
- Our point estimate for the true population difference in mean blood lead content is 15.97 mg.dl. The values in the interval (11.29, 20.65) mg/dl represent a reasonable range of estimates for the true population difference in mean blood lead content, and we are 90% confident that this method of creating a confidence interval will produce a result containing the true population mean difference.
- Were we to draw 100 samples of 33 matched pairs from the population described by this sample, and use each such sample to produce a confidence interval in this manner, approximately 90 of those confidence intervals would cover the true population mean difference in blood lead content levels.

18.9.2 Method 2

Or, we can apply the single-sample approach to a calculated difference in blood lead content between the exposed and control groups. Here, we'll get a **95%** two-sided confidence interval for that difference, instead of the 90% interval we obtained above.

```
tt2 <- t.test(bloodlead$exposed - bloodlead$control,  
conf.level = 0.95, alt = "two.sided")
```

```
tt2
```

One Sample t-test

```
data: bloodlead$exposed - bloodlead$control  
t = 5.783, df = 32, p-value = 2.036e-06  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
 10.34469 21.59470  
sample estimates:  
mean of x  
 15.9697
```

```
tidy(tt2) |> kable(digits = 2)
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
15.97	5.78	0	32	10.34	21.59	One Sample t-test	two.sided

18.9.3 Method 3

Or, we can provide R with two separate samples (unaffected and affected) and specify that the samples are paired. Here, we'll get a **99% one-sided** confidence interval (lower bound) for the population mean difference in blood lead content.

```
tt3 <- t.test(bloodlead$exposed, bloodlead$control, conf.level = 0.99,  
paired = TRUE, alt = "greater")
```

```
tt3
```

Paired t-test

```
data: bloodlead$exposed and bloodlead$control
```

```
t = 5.783, df = 32, p-value = 1.018e-06
alternative hypothesis: true mean difference is greater than 0
99 percent confidence interval:
 9.207658      Inf
sample estimates:
mean difference
 15.9697
```

```
tidy(tt3) |> kable(digits = 2)
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
15.97	5.78	0	32	9.21	Inf	Paired t-test	greater

Again, the three different methods using `t.test` for paired samples will all produce identical results if we feed them the same confidence level and type of interval (two-sided, greater than or less than).

18.9.4 Method 4

We can also use an intercept-only linear regression model to estimate the population mean of the paired differences with a two-tailed confidence interval, by creating a variable containing those paired differences.

```
model_lead <- lm(lead_diff ~ 1, data = bloodlead)

tidy(model_lead, conf.int = TRUE, conf.level = 0.95)

# A tibble: 1 x 7
  term       estimate std.error statistic   p.value conf.low conf.high
  <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept) 16.0      2.76      5.78 0.00000204    10.3     21.6
```

18.9.5 Method 5

If we have the data in a longer format, with a variable identifying the matched pairs, we can use a different specification for a linear model to obtain the same estimate.

```

model2_lead <- lm(lead_level ~ status + factor(pair), data = bloodlead_longer)

tidy(model2_lead, conf.int = TRUE, conf.level = 0.95) |>
  filter(term == "statusexposed")

# A tibble: 1 x 7
  term      estimate std.error statistic   p.value conf.low conf.high
  <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1 statusexposed    16.0      2.76     5.78 0.00000204     10.3     21.6

```

18.9.6 Assumptions

If we are building a confidence interval based on a sample of observations drawn from a population, then we must pay close attention to the assumptions of those procedures. The confidence interval procedure for the population mean paired difference using the t distribution assumes that:

1. We want to estimate the population mean paired difference.
2. We have drawn a sample of paired differences at random from the population of interest.
3. The sampled paired differences are drawn from the population set of paired differences independently and have identical distributions.
4. The population follows a Normal distribution. At the very least, the sample itself is approximately Normal.

18.10 Bootstrap CI for mean difference using paired samples

The same bootstrap approach is used for paired differences as for a single sample. We use the `smean.cl.boot()` function in the `Hmisc` package to obtain bootstrap confidence intervals for the population mean of the paired differences in blood lead content.

```

set.seed(431555)
Hmisc::smean.cl.boot(bloodlead$lead_diff, B = 1000, conf.int = 0.95)

```

Mean	Lower	Upper
15.96970	10.81742	21.48788

Note that in this case, the confidence interval for the difference in means is a bit less wide than the 95% confidence interval generated by the t test, which was (10.34, 21.59). It's common for the bootstrap to produce a narrower range (i.e. an apparently more precise estimate) for the

population mean, but it's not automatic that the endpoints from the bootstrap will be inside those provided by the t test, either.

For example, this bootstrap CI doesn't contain the t-test based interval, since its upper bound exceeds that of the t-based interval:

```
set.seed(431002)
Hmisc::smean.cl.boot(bloodlead$lead_diff, B = 1000, conf.int = 0.95)
```

Mean	Lower	Upper
15.96970	10.81667	21.66667

This demonstration aside, the appropriate thing to do when applying the bootstrap to specify a confidence interval is select a seed and the number ($B = 1,000$ or $10,000$, usually) of desired bootstrap replications, then run the bootstrap just once and move on, rather than repeating the process multiple times looking for a particular result.

18.10.1 Assumptions

The bootstrap confidence interval procedure for the population mean (or median) of a set of paired differences assumes that:

1. We want to estimate the population mean of the paired differences (or the population median).
2. We have drawn a sample of observations at random from the population of interest.
3. The sampled observations are drawn from the population of paired differences independently and have identical distributions.
4. We are willing to put up with the fact that different people (not using the same random seed) will get somewhat different confidence interval estimates using the same data.

As we've seen, a major part of the bootstrap's appeal is the ability to relax some assumptions.

18.11 Wilcoxon Signed Rank-based CI for paired samples

We could also use the Wilcoxon signed rank procedure to generate a CI for the pseudo-median of the paired differences.

```
wt <- wilcox.test(bloodlead$lead_diff, conf.int = TRUE,
                    conf.level = 0.90, exact = FALSE)
wt
```

```

Wilcoxon signed rank test with continuity correction

data: bloodlead$lead_diff
V = 499, p-value = 1.155e-05
alternative hypothesis: true location is not equal to 0
90 percent confidence interval:
10.99992 20.49998
sample estimates:
(pseudo)median
15.49996

```

```
tidy(wt)
```

```

# A tibble: 1 x 7
  estimate statistic  p.value conf.low conf.high method      alter~1
    <dbl>     <dbl>    <dbl>    <dbl>    <dbl>   <chr>      <chr>
1     15.5       499 0.0000115     11.0     20.5 Wilcoxon signed rank ~ two.si~
# ... with abbreviated variable name 1: alternative

```

As in the one sample case, we can revise this code slightly to specify a different confidence level, or gather a one-sided rather than a two-sided confidence interval.

18.11.1 Assumptions

The Wilcoxon signed rank confidence interval procedure in working with paired differences assumes that:

1. We want to estimate the population **pseudo-median** of the paired differences.
2. We have drawn a sample of observations at random from the population of paired differences of interest.
3. The sampled observations are drawn from the population of paired differences independently and have identical distributions.
4. The population follows a symmetric distribution. At the very least, the sample itself shows no substantial skew, so that the sample pseudo-median is a reasonable estimate for the population median.

18.12 Choosing a Confidence Interval Approach

Suppose we want to find a confidence interval for the population mean difference between two populations based on matched pairs.

1. If we are willing to assume that the population distribution is **Normal**
 - we usually use a t-based CI.
2. If we are **unwilling** to assume that the population is Normal,
 - use a **bootstrap** procedure to get a CI for the population mean, or even the median
 - but are willing to assume the population is symmetric, consider a **Wilcoxon signed rank** procedure to get a CI for the median, rather than the mean.

The two methods you'll use most often are the bootstrap (especially if the data don't appear to be at least pretty well fit by a Normal model) and the t-based confidence intervals (if the data do appear to fit a Normal model reasonably well.)

18.13 Conclusions for the bloodlead study

Using any of these procedures, we would conclude that the null hypothesis (that the true mean of the paired Exposed - Control differences is 0 mg/dl) is not consonant with what we see in the 90% confidence interval.

Procedure	Comparing	90% CI
Paired t	Means	11.3, 20.6
Wilcoxon signed rank	Pseudo-medians	11, 20.5
Bootstrap CI	Means	11.6, 20.6

Note that **one-sided** or **one-tailed** hypothesis testing procedures work the same way for paired samples as they did for a single sample.

18.14 The Sign test

The **sign test** is something we've skipped in our discussion so far. It is a test for consistent differences between pairs of observations, just as the paired t, Wilcoxon signed rank and bootstrap for paired samples can provide. It has the advantage that it is relatively easy to calculate by hand, and that it doesn't require the paired differences to follow a Normal distribution. In fact, it will even work if the data are substantially skewed.

- Calculate the paired difference for each pair, and drop those with difference = 0.
- Let N be the number of pairs that remain, so there are $2N$ data points.
- Let W , the test statistic, be the number of pairs (out of N) in which the difference is positive.
- Assuming that H_0 is true, then W follows a binomial distribution with probability 0.5 on N trials.

For example, consider our data on blood lead content:

```
bloodlead$lead_diff
```

```
[1] 22 5 23 -6 18 25 13 47 15 16 6 1 2 7 0 4 -9 -3 36 25 1 16 42 30 25
[26] 23 32 17 9 -3 60 14 14
```

Difference	# of Pairs
Greater than zero	28
Equal to zero	1
Less than zero	4

So we have $N = 32$ pairs, with $W = 28$ that are positive. We then use the `binom.test` approach in R:

```
binom.test(x = 28, n = 32, p = 0.5,
            alternative = "two.sided")
```

```
Exact binomial test

data: 28 and 32
number of successes = 28, number of trials = 32, p-value = 1.93e-05
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.7100516 0.9648693
sample estimates:
probability of success
 0.875
```

- A one-tailed test can be obtained by substituting in “less” or “greater” as the alternative of interest.
- The confidence interval provided doesn’t relate back to our original population means. It’s just showing the confidence interval around the probability of the exposed mean being greater than the control mean for a pair of children.

18.15 Paired (Dependent) vs. Independent Samples

One area that consistently trips students up in this course is the thought process involved in distinguishing studies comparing means that should be analyzed using *dependent* (i.e. paired or matched) samples and those which should be analyzed using *independent* samples. A dependent samples analysis uses additional information about the sample to pair/match subjects receiving the various exposures. That additional information is not part of an independent samples analysis (unpaired testing situation.) The reasons to do this are to (a) increase statistical power, and/or (b) reduce the effect of confounding. Here are a few thoughts on the subject.

In the design of experiments, **blocking** is the term often used for the process of arranging subjects into groups (blocks) that are similar to one another. Typically, a blocking factor is a source of variability that is not of primary interest to the researcher. An example of a blocking factor might be the sex of a patient; by blocking on sex, this source of variability is controlled for, thus leading to greater accuracy.

1. If the sample sizes are not balanced (not equal), the samples must be treated as independent, since there would be no way to precisely link all subjects. So, if we have 10 subjects receiving exposure A and 12 subjects receiving exposure B, a dependent samples analysis (such as a paired *t* test) is not correct.
2. The key element is a meaningful link between each observation in one exposure group and a specific observation in the other exposure group. Given a balanced design, the most common strategy indicating dependent samples involves two or more *repeated measures* on the same subjects. For example, if we are comparing outcomes *before* and *after* the application of an exposure, and we have, say, 20 subjects who provide us data both *before* and *after* the exposure, then the comparison of results *before* and *after* exposure should use a dependent samples analysis. The link between the subjects is the subject itself - each exposed subject serves as its own control.
3. The second most common strategy indicating dependent samples involves deliberate matching of subjects receiving the two exposures. A matched set of observations (often a pair, but it could be a trio or quartet, etc.) is determined using baseline information and then (if a pair is involved) one subject receives exposure A while the other member of the pair receives exposure B, so that by calculating the paired difference, we learn about the effect of the exposure, while controlling for the variables made similar across the two subjects by the matching process.
4. In order for a dependent samples analysis to be used, we need (a) a link between each observation across the exposure groups based on the way the data were collected, *and* (b) a consistent measure (with the same units of measurement) so that paired differences can be calculated and interpreted sensibly.

5. If the samples are collected to facilitate a dependent samples analysis, the correlation of the outcome measurements across the groups will often be moderately strong and positive. If that's the case, then the use of a dependent samples analysis will reduce the effect of baseline differences between the exposure groups, and thus provide a more precise estimate. But even if the correlation is quite small, a dependent samples analysis should provide a more powerful estimate of the impact of the exposure on the outcome than would an independent samples analysis with the same number of observations.

18.15.1 Three “Tricky” Examples

1. Suppose we take a convenient sample of 200 patients from the population of patients who complete a blood test in April 2017 including a check of triglycerides, and who have a triglyceride level in the high category (200 to 499 mg/dl). Next, we select a patient at random from this group of 200 patients, and then identify another patient from the group of 200 who is the same age (to within 2 years) and also the same sex. We then randomly assign our intervention to one of these two patients and usual care without our intervention to the other patient. We then set these two patients aside and return to our original sample, repeating the process until we cannot find any more patients in the same age range and of the same gender. This generates a total of 77 patients who receive the intervention and 77 who do not. If we are trying to assess the effect of our intervention on triglyceride level in October 2017 using this sample of 154 people, should we use dependent (paired) or independent samples?
2. Suppose we take a convenient sample of 77 patients from the population of patients who complete a blood test in April 2017 including a check of triglycerides, and who have a triglyceride level in the high category (200 to 499 mg/dl). Next, we take a convenient sample of 77 patients from the population of patients who complete a blood test in May 2017 including a check of triglycerides, and who have a triglyceride level in the high category (200 to 499 mg/dl). We flip a coin to determine whether the intervention will be given to each of the 77 patients from April 2017 (if the coin comes up “HEADS”) or instead to each of the 77 patients from May 2017 (if the coin comes up “TAILS”). Then, we assign our intervention to the patients seen in the month specified by the coin and assign usual care without our intervention to the patients seen in the other month. If we are trying to assess the effect of our intervention on triglyceride level in October 2017 using this sample of 154 people, should we use dependent (paired) or independent samples?
3. Suppose we take a convenient sample of 200 patients from the population of patients who complete a blood test in April 2017 including a check of triglycerides, and who have a triglyceride level in the high category (200 to 499 mg/dl). For each patient, we re-measure them again in October 2017, again checking their triglyceride level. But in between, we take the first 77 of the patients in a randomly sorted list and assign them to our intervention (which takes place from June through September 2017) and

take an additional group of 77 patients from the remaining part of the list and assign them to usual care without our intervention over the same time period. If we are trying to assess the effect of our intervention on each individual's change in triglyceride level (from April/May to October) using this sample of 154 people, should we use dependent (paired) or independent samples?

Answers to these “tricky” examples appear at the end of this Chapter.

18.16 A More Complete Decision Support Tool: Comparing Means

1. Are these paired or independent samples?
2. If paired samples, then are the paired differences approximately Normally distributed?
 - a. If yes, then a paired t test or confidence interval is likely the best choice.
 - b. If no, is the main concern outliers (with generally symmetric data), or skew?
 1. If the paired differences appear to be generally symmetric but with substantial outliers, a Wilcoxon signed rank test is an appropriate choice, as is a bootstrap confidence interval for the population mean of the paired differences.
 2. If the paired differences appear to be seriously skewed, then we'll usually build a bootstrap confidence interval, although a sign test is another reasonable possibility, although it doesn't provide a confidence interval for the population mean of the paired differences.
3. If independent, is each sample Normally distributed?
 - a. No → use Wilcoxon-Mann-Whitney rank sum test or bootstrap via `bootdif`.
 - b. Yes → are sample sizes equal?
 1. Balanced Design (equal sample sizes) - use pooled t test
 2. Unbalanced Design - use Welch test

18.16.1 Answers for the Three “Tricky” Examples

Answer for 1. Our first task is to identify the outcome and the exposure groups. Here, we are comparing the distribution of our outcome (triglyceride level in October) across two exposures: (a) receiving the intervention and (b) not receiving the intervention. We have a sample of 77 patients receiving the intervention, and a different sample of 77 patients receiving usual care. Each of the 77 subjects receiving the intervention is matched (on age and sex) to a specific subject not receiving the intervention. So, we can calculate paired differences by taking the triglyceride level for the exposed member of each pair and subtracting the triglyceride level for the usual care member of that same pair. Thus our comparison of the exposure groups should be accomplished using a *dependent* samples analysis, such as a paired t test.

Answer for 2. Again, we begin by identifying the outcome (triglyceride level in October) and the exposure groups. Here, we compare two exposures: (a) receiving the intervention and (b) receiving usual care. We have a sample of 77 patients receiving the intervention, and a different sample of 77 patients receiving usual care. But there is no pairing or matching involved. There is no connection implied by the way that the data were collected that implies that, for example, patient 1 in the intervention group is linked to any particular subject in the usual care group. So we need to analyze the data using independent samples.

Answer for 3. Once again, we identify the outcome (now it is the within-subject *change* in triglyceride level from April to October) and the exposure groups. Here again, we compare two exposures: (a) receiving the intervention and (b) receiving usual care. We have a sample of 77 patients receiving the intervention, and a different sample of 77 patients receiving usual care. But again, there is no pairing or matching between the patients receiving the intervention and the patients receiving usual care. While each outcome value is a difference (or change) in triglyceride levels, there's no connection implied by the way that the data were collected that implies that, for example, patient 1 in the intervention group is linked to any particular subject in the usual care group. So, again, we need to analyze the data using independent samples.

For more background and fundamental material, you might consider the Wikipedia pages on [Paired Difference Test](#) and on [Blocking \(statistics\)](#).

19 Hypothesis Testing: What is it good for?

19.1 Setup: Package Used Here

```
knitr::opts_chunk$set(comment = NA)

library(pwr)
```

19.2 Introduction

Hypothesis testing uses sample data to attempt to reject the hypothesis that nothing interesting is happening – that is, to reject the notion that chance alone can explain the sample results¹. We can, in many settings, use confidence intervals to summarize the results, as well, and confidence intervals and hypothesis tests are closely connected.

In particular, it's worth stressing that:

- **A significant effect is not necessarily the same thing as an interesting effect.** For example, results calculated from large samples are nearly always “significant” even when the effects are quite small in magnitude. Before doing a test, always ask if the effect is large enough to be of any practical interest. If not, why do the test?
- **A non-significant effect is not necessarily the same thing as no difference.** A large effect of real practical interest may still produce a non-significant result simply because the sample is too small.
- **There are assumptions behind all statistical inferences.** Checking assumptions is crucial to validating the inference made by any test or confidence interval.

¹Some of this is adapted from @GoodHardin, and @Utts1999

19.3 Five Steps in any Hypothesis Test

1. Specify the null hypothesis, H_0 (which usually indicates that there is no difference or no association between the results in various groups of subjects)
2. Specify the research or alternative hypothesis, H_A , sometimes called H_1 (which usually indicates that there is some difference or some association between the results in those same groups of subjects).
3. Specify the test procedure or test statistic to be used to make inferences to the population based on sample data. Here is where we usually specify α , the probability of incorrectly rejecting H_0 that we are willing to accept. In the absence of other information, we often use $\alpha = 0.05$
4. Obtain the data, and summarize it to obtain the relevant test statistic, which gets summarized as a p value.
5. Use the p value to either
 - **reject** H_0 in favor of the alternative H_A (concluding that there is a statistically significant difference/association at the α significance level) or
 - **retain** H_0 (and conclude that there is no statistically significant difference/association at the α significance level)

19.4 Type I and Type II Error

Once we know how unlikely the results would have been if the null hypothesis were true, we must make one of two choices:

1. The p value is not small enough to convincingly rule out chance. Therefore, we cannot reject the null hypothesis as an explanation for the results.
2. The p value was small enough to convincingly rule out chance. We reject the null hypothesis and accept the alternative hypothesis.

How small must the p value be in order to rule out the null hypothesis? The standard choice is 5%. This standardization has some substantial disadvantages². It is simply a convention that has become accepted over the years, and there are many situations for which a 5% cutoff is unwise. While it does give a specific level to keep in mind, it suggests a rather mindless cutpoint having nothing to do with the importance of the decision nor the costs or losses associated with outcomes.

²Ingelfinger JA, Mosteller F, Thibodeau LA and Ware JH (1987) Biostatistics in Clinical Medicine, 2nd Edition, New York: MacMillan. pp. 156-157.

19.5 The Courtroom Analogy

Consider the analogy of the jury in a courtroom.

1. The evidence is not strong enough to convincingly rule out that the defendant is innocent. Therefore, we cannot reject the null hypothesis, or innocence of the defendant.
2. The evidence was strong enough that we are willing to rule out the possibility that an innocent person (as stated in the null hypothesis) produced the observed data. We reject the null hypothesis, that the defendant is innocent, and assert the alternative hypothesis.

Consistent with our thinking in hypothesis testing, in many cases we would not accept the hypothesis that the defendant is innocent. We would simply conclude that the evidence was not strong enough to rule out the possibility of innocence.

The p value is the probability of getting a result as extreme or more extreme than the one observed if the proposed null hypothesis is true. Notice that it is not valid to actually accept that the null hypothesis is true. To do so would be to say that we are essentially convinced that chance alone produced the observed results – a common mistake.

19.6 Significance vs. Importance

Remember that a statistically significant relationship or difference does not necessarily mean an important one. A result that is significant in the statistical meaning of the word may not be significant clinically. Statistical significance is a technical term. Findings can be both statistically significant and practically significant or either or neither.

When we have large samples, we will regularly find small differences that have a small p value even though they have no practical importance. At the other extreme, with small samples, even large differences will often not be large enough to create a small p value. The notion of statistical significance has not helped science, and we won't perpetuate it any further.

19.7 What does Dr. Love dislike about p values and “statistical significance”?

A lot of things. A major issue is that I believe that p values are impossible to explain in a way that is both [a] technically correct and [b] straightforward at the same time. As evidence of this, you might want to look at [this article and associated video by Christie Aschwanden at 538.com](#)

The notion of a p value was an incredibly impressive achievement back when Wald and others were doing the work they were doing in the 1940s, and might still have been useful as recently

as 10 years ago. But the notion of a p value relies on a lot of flawed assumptions, and null hypothesis significance testing is fraught with difficulties. Nonetheless, researchers use p values every day.

19.8 The ASA Articles in 2016 and 2019 on Statistical Significance and P-Values

However, my primary motivation for taking the approach I'm taking comes from the pieces in two key reference collections we'll read and discuss more thoroughly in 431 and 432.

1. The American Statistical Association's 2016 [Statement on p-Values](#): Context, Process and Purpose.

The ASA Statement on p-Values and Statistical Significance (Wasserstein and Lazar 2016) was developed primarily because after decades, warnings about the don'ts had gone mostly unheeded. The statement was about what not to do, because there is widespread agreement about the don'ts.

2. [Statistical Inference in the 21st Century: A World Beyond \$p < 0.05\$](#) from 2019 in *The American Statistician*

This is a world where researchers are free to treat " $p = 0.051$ " and " $p = 0.049$ " as not being categorically different, where authors no longer find themselves constrained to selectively publish their results based on a single magic number. In this world, where studies with " $p < 0.05$ " and studies with " $p > 0.05$ " are not automatically in conflict, researchers will see their results more easily replicated—and, even when not, they will better understand why. As we venture down this path, we will begin to see fewer false alarms, fewer overlooked discoveries, and the development of more customized statistical strategies. Researchers will be free to communicate all their findings in all their glorious uncertainty, knowing their work is to be judged by the quality and effective communication of their science, and not by their p -values. As "statistical significance" is used less, statistical thinking will be used more. The ASA Statement on P-Values and Statistical Significance started moving us toward this world.... Now we must go further.

The ASA Statement on P-Values and Statistical Significance stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," " $p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.... Regardless of whether it was ever useful, a declaration of "statistical significance" has today become meaningless.

For the moment, I will say this. I emphasize confidence intervals over p values, which is at best a partial solution. But ...

1. Very rarely does a situation emerge in which a p value can be available in which looking at the associated confidence interval isn't far more helpful for making a comparison of interest.
2. The use of a p value requires making at least as many assumptions about the population, sample, individuals and data as does a confidence interval.
3. Most null hypotheses are clearly not exactly true prior to data collection, and so the test summarized by a p value is of questionable value most of the time.
4. No one has a truly adequate definition of a p value, in terms of both precision and parsimony. Brief, understandable definitions always fail to be technically accurate.
5. Bayesian approaches avoid some of these pitfalls, but come with their own issues.
6. Many smart people agree with me, and have sworn off of p values whenever they can.

Again, we'll look at these issues in greater depth later in the course.

19.9 Errors in Hypothesis Testing

In testing hypotheses, there are two potential decisions and each one brings with it the possibility that a mistake has been made.

Let's use the courtroom analogy. Here are the potential choices and associated potential errors. Although the seriousness of errors depends on the seriousness of the crime and punishment, the potential error for choice 2 is usually more serious.

1. We cannot rule out that the defendant is innocent, so (s)he is set free without penalty.
 - Potential Error: A criminal has been erroneously freed.
2. We believe that there is enough evidence to conclude that the defendant is guilty.
 - Potential Error: An innocent person is convicted / penalized and a guilty person remains free.

As another example, consider being tested for disease. Most tests for diseases are not 100% accurate. The lab technician or physician must make a choice:

1. In the opinion of the medical practitioner, you are healthy. The test result was weak enough to be called "negative" for the disease.
 - Potential Error: You actually have the disease but have been told that you do not. This is called a **false negative**.
2. In the opinion of the medical practitioner, you have the disease. The test results were strong enough to be called "positive" for the disease.

- Potential Error: You are actually healthy but have been told you have the disease. This is called a **false positive**.

19.10 The Two Types of Hypothesis Testing Errors

—	H_A is true	H_0 is true
Test Rejects H_0	Correct Decision	Type I Error (False Positive)
Test Retains H_0	Type II Error (False Negative)	Correct Decision

- A Type I error can only be made if the null hypothesis is actually true.
- A Type II error can only be made if the alternative hypothesis is actually true.

19.11 The Significance Level is the Probability of a Type I Error

If the null hypothesis is true, the p value is the probability of making an error by choosing the alternative hypothesis instead. Alpha (α) is defined as the probability of rejecting the null hypothesis when the null hypothesis is actually true, creating a Type I error. This is also called the significance level, so that $100(1-\alpha)$ is the confidence level – the probability of correctly concluding that there is no difference (retaining H_0) when the null hypothesis is true.

19.12 The Probability of avoiding a Type II Error is called Power

A Type II error is made if the alternative hypothesis is true, but you fail to choose it. The probability depends on exactly which part of the alternative hypothesis is true, so that computing the probability of making a Type II error is not feasible. The power of a test is the probability of making the correct decision when the alternative hypothesis is true. Beta (β) is defined as the probability of concluding that there was no difference, when in fact there was one (a Type II error). Power is then just $1 - \beta$, the probability of concluding that there was a difference, when, in fact, there was one.

Traditionally, people like the power of a test to be at least 80%, meaning that β is at most 0.20. Often, I'll be arguing for 90% as a minimum power requirement, or we'll be presenting a range of power calculations for a variety of sample size choices.

19.13 Incorporating the Costs of Various Types of Errors

Which error is more serious in medical testing, where we think of our H_0 : patient is healthy vs. H_A : disease is present?

It depends on the disease and on the consequences of a negative or positive test result. A false negative in a screening test for cancer could lead to a fatal delay in treatment, whereas a false positive would probably lead to a retest. A more troublesome example occurs in testing for an infectious disease. Inevitably, there is a trade-off between the two types of errors. It all depends on the consequences.

It would be nice if we could specify the probability that we were making an error with each potential decision. We could then weigh the consequence of the error against its probability. Unfortunately, in most cases, we can only specify the conditional probability of making a Type I error, given that the null hypothesis is true.

In deciding whether to reject a null hypothesis, we will need to consider the consequences of the two potential types of errors. If a Type I error is very serious, then you should reject the null hypothesis only if the p value is very small. Conversely, if a Type II error is more serious, you should be willing to reject the null hypothesis with a larger p value, perhaps 0.10 or 0.20, instead of 0.05.

- α is the probability of rejecting H_0 when H_0 is true.
 - So $1 - \alpha$, the confidence level, is the probability of retaining H_0 when that's the right thing to do.
- β is the probability of retaining H_0 when H_A is true.
 - So $1 - \beta$, the power, is the probability of rejecting H_0 when that's the right thing to do.

–	H_A is True	H_0 is True
Test Re- jects H_0	Correct Decision ($1 - \beta$)	Type I Error (α)
Test Re- tains H_0	Type II Error (β)	Correct Decision ($1 - \alpha$)

19.14 Power and Sample Size Considerations

For most statistical tests, it is theoretically possible to estimate the power of the test in the design stage, (before any data are collected) for various sample sizes, so we can hone in on a sample size choice which will enable us to collect data only on as many subjects as are truly necessary.

A power calculation is likely the most common element of a scientific grant proposal on which a statistician is consulted. This is a fine idea in theory, but in practice...

- The tests that have power calculations worked out in intensive detail using R are mostly those with more substantial assumptions. Examples include t tests that assume population normality, common population variance and balanced designs in the independent samples setting, or paired t tests that assume population normality in the paired samples setting.
- These power calculations are also usually based on tests rather than confidence intervals, which would be much more useful in most settings. Simulation is your friend here.
- Even more unfortunately, this process of doing power and related calculations is **far more of an art than a science**.
- As a result, the value of many power calculations is negligible, since the assumptions being made are so arbitrary and poorly connected to real data.
- On several occasions, I have stood in front of a large audience of medical statisticians actively engaged in clinical trials and other studies that require power calculations for funding. When I ask for a show of hands of people who have had power calculations prior to such a study whose assumptions matched the eventual data perfectly, I get lots of laughs. It doesn't happen.
- Even the underlying framework that assumes a power of 80% with a significance level of 5% is sufficient for most studies is pretty silly.

All that said, I feel obliged to show you some examples of power calculations done using R, and provide some insight on how to make some of the key assumptions in a way that won't alert reviewers too much to the silliness of the enterprise. All of the situations described in this Chapter are toy problems, but they may be instructive about some fundamental ideas.

19.15 Sample Size in a One-Sample t test

For a t test, R can estimate any one of the following elements, given the other four, using the `power.t.test` command, for either a one-tailed or two-tailed single-sample t test...

- n = the sample size
- δ = delta = the true difference in population means between the null hypothesis value and a particular alternative

- $s = \text{sd}$ = the true standard deviation of the population
- $\alpha = \text{sig.level}$ = the significance level for the test (maximum acceptable risk of Type I error)
- $1 - \beta = \text{power}$ = the power of the t test to detect the effect of size δ

19.15.1 A Toy Example

Suppose that in a recent health survey, the average beef consumption in the U.S. per person was 90 pounds per year. Suppose you are planning a new study to see if beef consumption levels have changed. You plan to take a random sample of 25 people to build your new estimate, and test whether the current pounds of beef consumed per year is 90. Suppose you want to do a two-sided (two-tailed) test at 95% confidence (so $\alpha = 0.05$), and that you expect that the true difference will need to be at least $\delta = 5$ pounds (i.e. 85 or less or 95 or more) in order for the result to be of any real, practical interest. Suppose also that you are willing to assume that the true standard deviation of the measurements in the population is 10 pounds.

That is, of course, a lot to suppose.

Now, we want to know what power the proposed experiment will have to detect a change of 5 pounds (or more) away from the original 90 pounds, with these specifications, and how tweaking these specifications will affect the power of the study.

So, we have - $n = 25$ data points to be collected - $\delta = 5$ pounds is the minimum clinically meaningful effect size - $s = 10$ is the assumed population standard deviation, in pounds per year - α is 0.05, and we'll do a two-sided test

19.15.2 Using the `power.t.test` function

```
power.t.test(n = 25, delta = 5, sd = 10, sig.level = 0.05,
             type="one.sample", alternative="two.sided")
```

One-sample t test power calculation

```
n = 25
delta = 5
sd = 10
sig.level = 0.05
power = 0.6697014
alternative = two.sided
```

So, under this study design, we would expect to detect an effect of size $\delta = 5$ pounds with just under 67% power, i.e. with a probability of incorrect retention of H_0 of just about 1/3. Most of the time, we'd like to improve this power, and to do so, we'd need to adjust our assumptions.

19.16 Changing Assumptions

We made assumptions about the sample size n , the minimum clinically meaningful effect size (change in the population mean) δ , the population standard deviation s , and the significance level α , not to mention decisions about the test, like that we'd do a one-sample t test, rather than another sort of test for a single sample, and that we'd do a two-tailed, or two-sided test. Often, these assumptions are tweaked a bit to make the power look more like what a reviewer/funder is hoping to see.

19.16.1 Increasing Sample Size Increases Power

Suppose, we committed to using more resources and gathering data from 40 subjects instead of the 25 we assumed initially – what effect would this have on our power?

```
power.t.test(n = 40, delta = 5, sd = 10, sig.level = 0.05,
              type="one.sample", alternative="two.sided")
```

```
One-sample t test power calculation
```

```
  n = 40
  delta = 5
  sd = 10
  sig.level = 0.05
  power = 0.8693979
  alternative = two.sided
```

With more samples, we should have a more powerful test, able to detect the difference with greater probability. In fact, a sample of 40 paired differences yields 87% power. As it turns out, we would need at least 44 observations with this scenario to get to 90% power, as shown in the calculation below, which puts the power in, but leaves out the sample size.

```
power.t.test(power=0.9, delta = 5, sd = 10, sig.level = 0.05,
              type="one.sample", alternative="two.sided")
```

```
One-sample t test power calculation
```

```
n = 43.99552
delta = 5
sd = 10
sig.level = 0.05
power = 0.9
alternative = two.sided
```

We see that we would need at least 44 observations to achieve 90% power. Note: we always round the sample size up in doing a power calculation – if this calculation had actually suggested $n = 43.1$ paired differences were needed, we would still have rounded up to 44.

19.16.2 Increasing Effect Size will increase Power

A larger effect should be easier to detect. If we go back to our original calculation, which had 67% power to detect an effect of size $\delta = 5$, and now change the desired effect size to $\delta = 6$ pounds (i.e. a value of 84 or less or 96 or more), we should obtain a more powerful design.

```
power.t.test(n = 25, delta = 6, sd = 10, sig.level = 0.05,
              type="one.sample", alternative="two.sided")
```

```
One-sample t test power calculation
```

```
n = 25
delta = 6
sd = 10
sig.level = 0.05
power = 0.8207213
alternative = two.sided
```

We see that this change in effect size from 5 to 6, leaving everything else the same, increases our power from 67% to 82%. To reach 90% power, we'd need to increase the effect size we were trying to detect to at least 6.76 pounds.

```
power.t.test(n = 25, power = 0.9, sd = 10, sig.level = 0.05,
              type="one.sample", alternative="two.sided")
```

```
One-sample t test power calculation
```

```
n = 25
delta = 6.759051
sd = 10
sig.level = 0.05
power = 0.9
alternative = two.sided
```

- Again, note that I am rounding up here.
- Using $\delta = 6.75$ would not quite make it to 90.00% power.
- Using $\delta = 6.76$ guarantees that the power will be 90% or more, and not just round up to 90%.

19.16.3 Decreasing the Standard Deviation will increase Power

The choice of standard deviation is usually motivated by a pilot study, or else pulled out of thin air - it's relatively easy to convince yourself that the true standard deviation might be a little smaller than you'd guessed initially. Let's see what happens to the power if we reduce the sample standard deviation from 10 pounds to 9. This should make the effect of 5 pounds easier to detect, because it will have smaller variation associated with it.

```
power.t.test(n = 25, delta = 5, sd = 9, sig.level = 0.05,
              type="one.sample", alternative="two.sided")
```

```
One-sample t test power calculation
```

```
n = 25
delta = 5
sd = 9
sig.level = 0.05
power = 0.759672
alternative = two.sided
```

This change in standard deviation from 10 to 9, leaving everything else the same, increases our power from 67% to nearly 76%. To reach 90% power, we'd need to decrease the standard deviation of the population paired differences to no more than 7.39 pounds.

```
power.t.test(n = 25, delta = 5, sd = NULL, power = 0.9, sig.level = 0.05,
             type="one.sample", alternative="two.sided")
```

```
One-sample t test power calculation
```

```
  n = 25
  delta = 5
  sd = 7.397486
  sig.level = 0.05
  power = 0.9
  alternative = two.sided
```

Note I am rounding down here.

- Using $s = 7.4$ pounds would not quite make it to 90.00% power.

Note also that in order to get R to treat the standard deviation as unknown, I must specify it as `NULL` in the formula.

19.16.4 Larger Significance Level increases Power

We can trade off some of our Type II error (lack of power) for Type I error. If we are willing to trade off some Type I error (as described by the α), we can improve the power. For instance, suppose we decided to run the original test with 90% confidence.

```
power.t.test(n = 25, delta = 5, sd = 10, sig.level = 0.1,
             type="one.sample", alternative="two.sided")
```

```
One-sample t test power calculation
```

```
  n = 25
  delta = 5
  sd = 10
  sig.level = 0.1
  power = 0.7833861
  alternative = two.sided
```

The calculation suggests that our power would thus increase from 67% to just over 78%.

19.17 Paired Sample t Tests and Power/Sample Size

For a paired-samples t test, R can estimate any one of the following elements, given the other four, using the `power.t.test` command, for either a one-tailed or two-tailed paired t test...

- n = the sample size (# of pairs) being compared
- δ = delta = the true difference in means between the two groups
- $s = sd$ = the true standard deviation of the paired differences
- $\alpha = \text{sig.level}$ = the significance level for the comparison (maximum acceptable risk of Type I error)
- $1 - \beta$ = power = the power of the paired t test to detect the effect of size δ

19.17.1 A Toy Example

As a toy example, suppose you are planning a paired samples experiment involving $n = 30$ subjects who will each provide a “Before” and an “After” result, which is measured in days.

Suppose you want to do a two-sided (two-tailed) test at 95% confidence (so $\alpha = 0.05$), and that you expect that the true difference between the “Before” and “After” groups will have to be at least $\delta = 5$ days to be of any real interest. Suppose also that you are willing to assume that the true standard deviation of those paired differences will be 10 days.

That is, of course, a lot to suppose.

Now, we want to know what power the proposed experiment will have to detect this difference with these specifications, and how tweaking these specifications will affect the power of the study.

So, we have - $n = 30$ paired differences will be collected - $\delta = 5$ days is the minimum clinically meaningful difference - $s = 10$ days is the assumed population standard deviation of the paired differences - α is 0.05, and we'll do a two-sided test

19.17.2 Using the `power.t.test` function

```
power.t.test(n = 30, delta = 5, sd = 10, sig.level = 0.05,
             type="paired", alternative="two.sided")
```

```
Paired t test power calculation
```

```
n = 30
delta = 5
```

```
sd = 10
sig.level = 0.05
power = 0.7539627
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

So, under this study design, we would expect to detect an effect of size $\delta = 5$ days with 75% power, i.e. with a probability of incorrect retention of H_0 of 0.25. Most of the time, we'd like to improve this power, and to do so, we'd need to adjust our assumptions.

19.17.3 Changing Assumptions in a Power Calculation

We made assumptions about the sample size n, the minimum clinically meaningful difference in means δ , the population standard deviation s, and the significance level α , not to mention decisions about the test, like that we'd do a paired t test, rather than another sort of test for paired samples, or use an independent samples approach, and that we'd do a two-tailed, or two-sided test. Often, these assumptions are tweaked a bit to make the power look more like what a reviewer/funder is hoping to see.

19.17.4 Changing the Sample Size

Suppose, we committed to using more resources and gathering “Before” and “After” data from 40 subjects instead of the 30 we assumed initially – what effect would this have on our power?

```
power.t.test(n = 40, delta = 5, sd = 10, sig.level = 0.05,
              type="paired", alternative="two.sided")
```

Paired t test power calculation

```
n = 40
delta = 5
sd = 10
sig.level = 0.05
power = 0.8693979
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

With more samples, we should have a more powerful test, able to detect the difference with greater probability. In fact, a sample of 40 paired differences yields 87% power. As it turns out, we would need at least 44 paired differences with this scenario to get to 90% power, as shown in the calculation below, which puts the power in, but leaves out the sample size.

```
power.t.test(power=0.9, delta = 5, sd = 10, sig.level = 0.05,
             type="paired", alternative="two.sided")
```

Paired t test power calculation

```
n = 43.99552
delta = 5
sd = 10
sig.level = 0.05
power = 0.9
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

We see that we would need at least 44 paired differences to achieve 90% power. Note: we always round the sample size up in doing a power calculation – if this calculation had actually suggested $n = 43.1$ paired differences were needed, we would still have rounded up to 44.

19.17.5 Changing the Effect Size

A larger effect should be easier to detect. If we go back to our original calculation, which had 75% power to detect an effect (i.e. a true population mean difference) of size $\delta = 5$, and now change the desired effect size to $\delta = 6$, we should obtain a more powerful design.

```
power.t.test(n = 30, delta = 6, sd = 10, sig.level = 0.05,
             type="paired", alternative="two.sided")
```

Paired t test power calculation

```
n = 30
delta = 6
sd = 10
sig.level = 0.05
```

```
power = 0.887962
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

We see that this change in effect size from 5 to 6, leaving everything else the same, increases our power from 75% to nearly 89%. To reach 90% power, we'd need to increase the effect size we were trying to detect to at least 6.13 days.

- Again, note that I am rounding up here.
- Using $\delta = 6.12$ would not quite make it to 90.00% power.
- Using $\delta = 6.13$ guarantees that the power will be 90% or more, and not just round up to 90%..

19.17.6 Changing the Standard Deviation

The choice of standard deviation is usually motivated by a pilot study, or else pulled out of thin air. It's relatively easy to convince yourself that the true standard deviation might be a little smaller than you'd guessed initially. Let's see what happens to the power if we reduce the sample standard deviation from 10 days to 9 days. This should make the effect of 5 days easier to detect as being different from the null hypothesized value of 0, because it will have smaller variation associated with it.

```
power.t.test(n = 30, delta = 5, sd = 9, sig.level = 0.05,
              type="paired", alternative="two.sided")
```

Paired t test power calculation

```
n = 30
delta = 5
sd = 9
sig.level = 0.05
power = 0.8366514
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

This change in standard deviation from 10 to 9, leaving everything else the same, increases our power from 75% to nearly 84%. To reach 90% power, we'd need to decrease the standard deviation of the population paired differences to no more than 8.16 days.

Note I am rounding down here, because using $s = 8.17$ days would not quite make it to 90.00% power. Note also that in order to get R to treat the sd as unknown, I must specify it as NULL in the formula...

```
power.t.test(n = 30, delta = 5, sd = NULL, power = 0.9,
             sig.level = 0.05, type="paired", alternative="two.sided")
```

Paired t test power calculation

```
n = 30
delta = 5
sd = 8.163989
sig.level = 0.05
power = 0.9
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

19.17.7 Changing the Significance Level

We can trade off some of our Type II error (lack of power) for Type I error. If we are willing to trade off some Type I error (as described by the α), we can improve the power. For instance, suppose we decided to run the original test with 90% confidence.

```
power.t.test(n = 30, delta = 5, sd = 10, sig.level = 0.1,
             type="paired", alternative="two.sided")
```

Paired t test power calculation

```
n = 30
delta = 5
sd = 10
sig.level = 0.1
power = 0.8482542
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

The calculation suggests that our power would thus increase from 75% to nearly 85%.

19.18 Two Independent Samples: Power for t Tests

For an independent-samples t test, with a balanced design (so that $n_1 = n_2$), R can estimate any one of the following elements, given the other four, using the `power.t.test` command, for either a one-tailed or two-tailed t test...

- n = the sample size in each of the two groups being compared
- δ = delta = the true difference in means between the two groups
- $s = sd$ = the true standard deviation of the individual values in each group (assumed to be constant – since we assume equal population variances)
- $\alpha = \text{sig.level}$ = the significance level for the comparison (maximum acceptable risk of Type I error)
- $1 - \beta = \text{power}$ = the power of the t test to detect the effect of size δ

This method only produces power calculations for balanced designs – where the sample size is equal in the two groups. If you want a two-sample power calculation for an unbalanced design, you will need to use a different library and function in R, as we'll see.

19.19 A New Example

Suppose we plan a study of the time to relapse for patients in a drug trial, where subjects will be assigned randomly to a (new) treatment or to a placebo. Suppose we anticipate that the placebo group will have a mean of about 9 months, and want to detect an improvement (increase) in time to relapse of 50%, so that the treatment group would have a mean of at least 13.5 months. We'll use $\alpha = .10$ and $\beta = .10$, as well. Assume we'd do a two-sided test, with an equal number of observations in each group, and we'll assume the observed standard deviation of 9 months in a pilot study will hold here, as well.

We want the sample size required by the test under a two sample setting where:

- $\alpha = .10$,
 - with 90% power (so that $\beta = .10$),
 - and where we will have equal numbers of samples in the placebo group (group 1) and the treatment group (group 2).
-
- We'll plug in the observed standard deviation of 9 months.
 - We'll look at detecting a change from 9 [the average in the placebo group] to 13.5 (a difference of 50%, giving delta = 4.5)
 - using a two-sided pooled t-test.

The appropriate R command is:

```
power.t.test(delta = 4.5, sd = 9,
             sig.level = 0.10, power = 0.9,
             type="two.sample",
             alternative="two.sided")
```

Two-sample t test power calculation

```
n = 69.19782
delta = 4.5
sd = 9
sig.level = 0.1
power = 0.9
alternative = two.sided
```

NOTE: n is number in *each* group

This suggests that we will need a sample of at least 70 subjects in the treated group and an additional 70 subjects in the placebo group, for a total of 140 subjects.

19.19.1 Another Scenario

What if resources are sparse, and we'll be forced to do the study with no more than 120 subjects, overall? If we require 90% confidence in a two-sided test, what power will we have?

```
power.t.test(n = 60, delta = 4.5, sd = 9,
             sig.level = 0.10,
             type="two.sample",
             alternative="two.sided")
```

Two-sample t test power calculation

```
n = 60
delta = 4.5
sd = 9
sig.level = 0.1
power = 0.859484
alternative = two.sided
```

NOTE: n is number in *each* group

It looks like the power under those circumstances would be just under 86%. Note that the $n = 60$ refers to half of the total sample size, since we'll need 60 drug and 60 placebo subjects in this balanced design.

19.20 Power for Independent Sample T tests with Unbalanced Designs

Using the `pwr` package, R can do sample size calculations that describe the power of a two-sample t test that does not require a balanced design using the `pwr.t2n.test` command.

Suppose we wanted to do the same study as we described above, using 100 “treated” patients but as few “placebo” patients as possible. What sample size would be required to maintain 90% power? There is one change here – the effect size d in the `pwr.t2n.test` command is specified using the difference in means δ that we used previously, divided by the standard deviation s that we used previously. So, in our old setup, we assumed $\delta = 4.5$, $s = 9$, so now we'll assume $d = 4.5/9$ instead.

```
pwr.t2n.test(n1 = 100, d = 4.5/9,
              sig.level = 0.1, power = 0.9,
              alternative="two.sided")
```

```
t test power calculation

n1 = 100
n2 = 52.82433
d = 0.5
sig.level = 0.1
power = 0.9
alternative = two.sided
```

We would need at least 53 subjects in the “placebo” group.

19.20.1 The most efficient design for an independent samples comparison will be balanced.

- Note that if we use $n_1 = 100$ subjects in the treated group, we need at least $n_2 = 53$ in the placebo group to achieve 90% power, and a total of 153 subjects.
- Compare this to the balanced design, where we needed 70 subjects in each group to achieve the same power, thus, a total of 140 subjects.

We saw earlier that a test with 60 subjects in each group would yield just under 86% power. Suppose we instead built a test with 80 subjects in the treated group, and 40 in the placebo group, then what would our power be?

```
pwr.t2n.test(n1 = 80, n2 = 40, d = 4.5/9,  
             sig.level = 0.10,  
             alternative="two.sided")
```

```
t test power calculation  
  
n1 = 80  
n2 = 40  
d = 0.5  
sig.level = 0.1  
power = 0.821823  
alternative = two.sided
```

As we'd expect, the power is stronger for a balanced design than for an unbalanced design with the same overall sample size.

Note that I used a two-sided test to establish my power calculation – in general, this is the most conservative and defensible approach for any such calculation, **unless there is a strong and specific reason to use a one-sided approach in building a power calculation, don't.**

20 Two Examples Comparing Means

20.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

source("data/Love-boost.R")
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

In addition to the `Love-boost.R` script, we will also use the `favstats` function from the `mosaic` package.

20.2 A Study of Battery Life

Should you buy generic rather than brand-name batteries? Bock, Velleman, and De Veaux (2004) describe a designed experiment to test battery life. A (male) student obtained six pairs of AA alkaline batteries from two major battery manufacturers; a well-known brand name and a generic brand, so that battery brand was the factor of interest.

To estimate the difference in mean lifetimes across the two manufacturers, the student kept a battery-powered CD player with the same CD running continuously, with the volume control fixed at 5, and measured the time until no more music was heard through the headphones. (He ran an initial trial to find out approximately how long that would take, so he didn't have to spend the first 3 hours of each run listening to the same CD.) The outcome was the time in minutes until the sound stopped. To account for changes in the CD player's performance over time, he randomized the run order by choosing pairs of batteries (the CD-player required two batteries to run) at random.

Here are the results for the 6 brand name and 6 generic tests, in minutes, found in the `battery.csv` data file, where `run` indicates the order in which the tests were run...

```

battery <- read_csv("data/battery.csv",
                     show_col_types = FALSE)

battery

# A tibble: 12 x 4
  run test type     time
  <dbl> <dbl> <chr>   <dbl>
1     1     1 brand name 191.
2     2     2 brand name 206.
3     6     3 brand name 199.
4     8     4 brand name 172.
5     9     5 brand name 184
6    12     6 brand name 170.
7     3     1 generic   194
8     4     2 generic   204.
9     5     3 generic   204.
10    7     4 generic   206.
11    10    5 generic   222.
12    11    6 generic   209.

```

20.2.1 Question 1. What is the outcome under study?

We are studying battery lifetimes (time until the sound stopped) in minutes.

20.2.2 Question 2. What are the treatment/exposure groups?

We are comparing the two brands of batteries: the well-known vs. the generic.

20.2.3 Question 3. Are the data collected using paired or independent samples?

Of course, if we had different numbers of samples in the two groups, then we'd know without further thought that independent samples were required. Since we have 6 observations in the brand name group, and also have 6 observations in the generic group, i.e. a balanced design, we need to pause now to decide whether paired or independent samples testing is appropriate in this setting.

Two samples are paired if each data point in one sample is naturally linked to a specific data point in the other sample. So, do we have paired or independent samples?

- Despite the way I've set up the data table, there is no particular reason to pair, say, run #1 (a brand name run) with any particular experimental run in the generic group. So the samples are independent. This is not a matched-pairs design.
- In each trial, the student either used two of the well-known batteries, or two of the generic batteries.
- Any of the tests/confidence intervals for the independent samples methods suggests a statistically significant (at the 5% level) difference between the generic and brand name batteries.

20.2.4 Question 4. Are the data a random sample from the population of interest?

Probably not. The data are likely to come from a convenient sample of batteries. I don't know how this might bias the study, though. It seems unlikely that there would be a particular bias unless, for example, the well-known batteries were substantially older or younger than the generic.

20.2.5 Question 5. What significance level will we use?

We have no reason not to use a 95% confidence level.

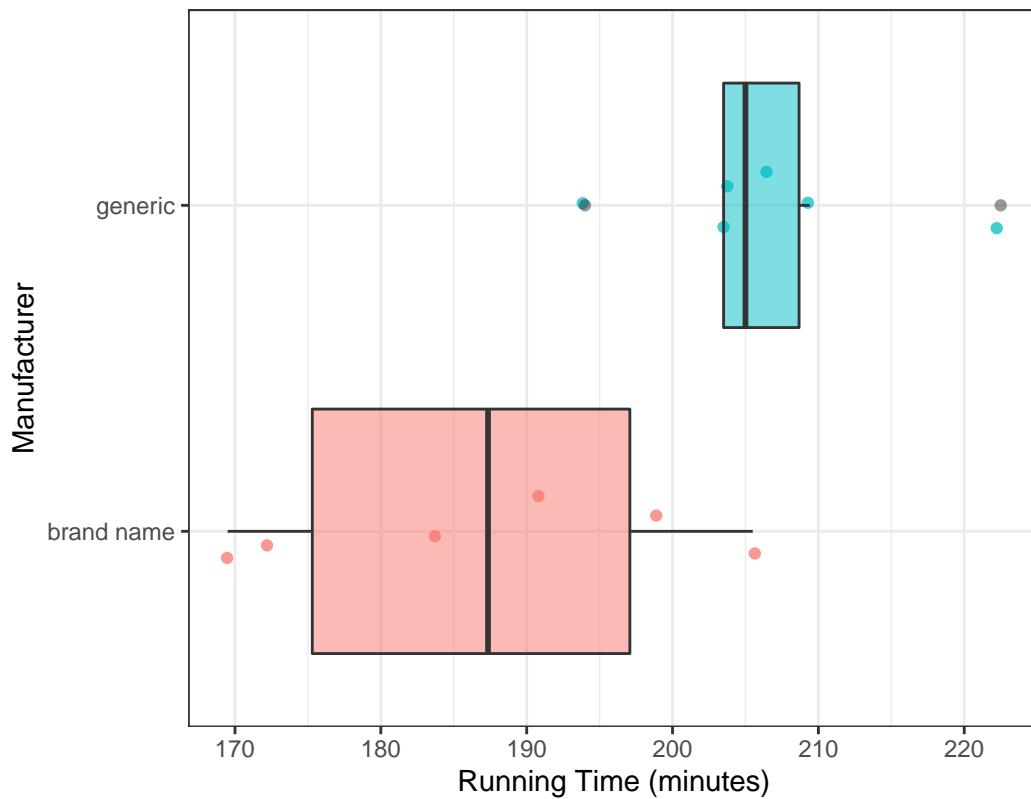
20.2.6 Question 6. Are we using a one-sided or two-sided comparison?

We could argue for a one-sided comparison, but I'll be safe and use the two-sided version.

20.2.7 Question 9. What does the distribution of outcomes in each group tell us?

```
ggplot(battery, aes(x = type, y = time, fill = type)) +
  geom_jitter(aes(color = type), alpha = 0.75, width = 0.125) +
  geom_boxplot(alpha = 0.5) +
  coord_flip() +
  guides(fill = "none", col = "none") +
  labs(title = "Battery Running Time, by Manufacturer",
       y = "Running Time (minutes)", x = "Manufacturer")
```

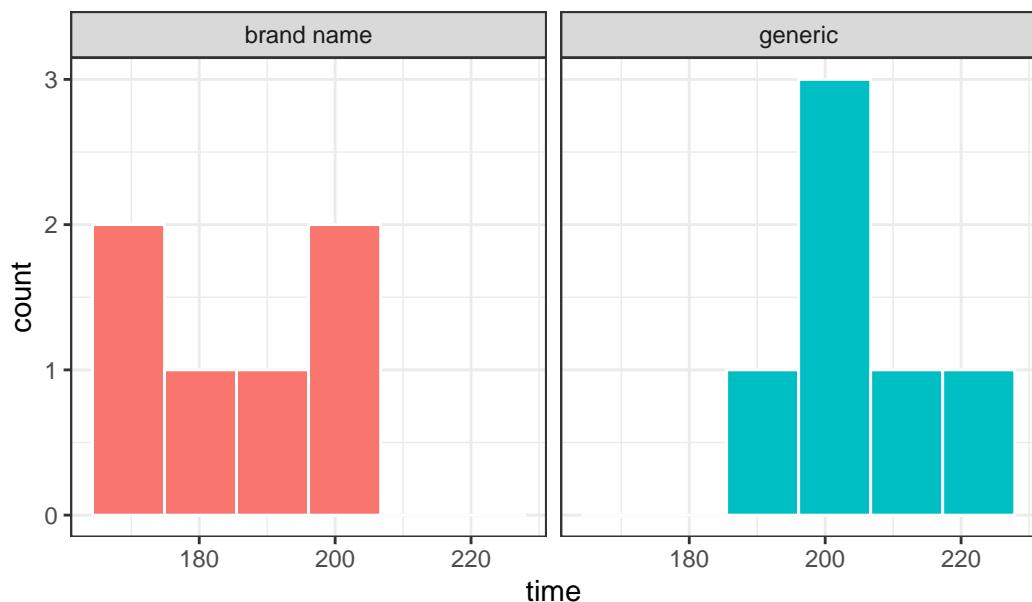
Battery Running Time, by Manufacturer



We can generate histograms, too, but that's an issue, because we have so few observations.

```
ggplot(battery, aes(x = time, fill = type)) +  
  geom_histogram(bins = 6, col = "white") +  
  facet_wrap(~ type) +  
  guides(fill = "none") +  
  labs(title = "Battery Running Time, by Manufacturer")
```

Battery Running Time, by Manufacturer



```
mosaic::favstats(time ~ type, data = battery)
```

```
Registered S3 method overwritten by 'mosaic':  
method                  from  
fortify.SpatialPolygonsDataFrame ggplot2
```

	type	min	Q1	median	Q3	max	mean	sd	n	missing
1	brand name	169.5	175.3	187.35	197.075	205.5	186.8833	14.374341	6	0
2	generic	194.0	203.5	205.00	208.675	222.5	206.5667	9.366251	6	0

It sure looks like the generic batteries lasted longer. And they also look like they were more consistent. The sample means are 206.6 for the generic group, 186.9 minutes for brand name, so the point estimate of the difference is 19.7 minutes.

The question is: can we be confident that the difference we observe here is more than just random fluctuation, at a 5% significance level?

20.2.8 Inferential Results for the Battery Study

In the table below, I have summarized the two-sided testing results for most of the ways in which we have looked at a two sample comparison so far, with 95% confidence intervals. If the

samples really are paired, then we must choose from the paired samples comparisons described in the table. If the samples really are independent, then we must choose from the independent samples comparisons.

20.2.9 Paired Samples Approaches

Method	<i>p</i> Value	95% CI for Generic - Brand Name
Paired t	0.058	-1.0, 40.4
Wilcoxon signed rank	0.063	-2.0, 39.9
Bootstrap via <code>smean.cl.boot</code>	-	6.7, 33.0

20.2.10 Independent Samples Approaches

Method	<i>p</i> Value	95% CI for Generic - Brand Name
Pooled t	0.018	4.1, 35.3
Welch's t	0.021	3.7, 35.6
Wilcoxon Mann Whitney rank sum	0.030	3.3, 37.0
Bootstrap via <code>bootdif</code>	-	7.7, 32.2

20.3 The Breakfast Study: Does Oat Bran Cereal Lower Serum LDL Cholesterol?

Norman and Streiner (2014) describe a crossover study that was conducted to investigate whether oat bran cereal helps to lower serum cholesterol levels in hypercholesterolemic males. Fourteen such individuals were randomly placed on a diet that included either oat bran or corn flakes; after two weeks, their low-density lipoprotein (LDL) cholesterol levels, in mmol/l were recorded. Each subject was then switched to the alternative diet. After a second two-week period, the LDL cholesterol level of each subject was again recorded.

```
breakfast <- read_csv("data/breakfast.csv",
                      show_col_types = FALSE)

breakfast

# A tibble: 14 x 3
  subject cornflakes oatbran
  <fct>    <dbl>     <dbl>
1 cornflakes  10.0      10.0
2 cornflakes  10.0      10.0
3 cornflakes  10.0      10.0
4 cornflakes  10.0      10.0
5 cornflakes  10.0      10.0
6 cornflakes  10.0      10.0
7 cornflakes  10.0      10.0
8 cornflakes  10.0      10.0
9 cornflakes  10.0      10.0
10 cornflakes 10.0      10.0
11 cornflakes 10.0      10.0
12 cornflakes 10.0      10.0
13 cornflakes 10.0      10.0
14 cornflakes 10.0      10.0
```

	<dbl>	<dbl>	<dbl>
1	1	4.61	3.84
2	2	6.42	5.57
3	3	5.4	5.85
4	4	4.54	4.8
5	5	3.98	3.68
6	6	3.82	2.96
7	7	5.01	4.41
8	8	4.34	3.72
9	9	3.8	3.49
10	10	4.56	3.84
11	11	5.35	5.26
12	12	3.89	3.73
13	13	2.25	1.84
14	14	4.24	4.14

20.3.1 Question 1. What is the outcome under study?

We are studying levels of LDL cholesterol, in mmol/l. Note that if we wanted to convert to a more familiar scale, specifically mg/dl, we would multiply the mmol/l by 18, as it turns out.

20.3.2 Question 2. What are the treatment/exposure groups?

We are comparing subjects after two weeks of eating corn flakes to the same subjects after two weeks of eating oat bran.

20.3.3 Question 3. Are the data collected using paired or independent samples?

These are matched pairs, paired by subject. Each subject produced an oat bran result and a corn flakes result.

20.3.4 Question 4. Are the data a random sample from the population of interest?

Probably not. The data are likely to come from a convenient sample of 14 individuals but they were randomly assigned to cornflakes first or to oat bran first, then crossed over.

20.3.5 Question 5. What significance level will we use?

We have no reason not to use our usual 95% confidence level, so `alpha = 0.05`

20.3.6 Question 6. Are we using a one-sided or two-sided comparison?

We could argue for a one-sided comparison, but I'll be safe and use the two-sided version.

20.3.7 Question 7. Did pairing help reduce nuisance variation?

After we drop the `breakfast.csv` file into the `breakfast` data frame, we look at the correlation of cornflakes and oatbran results across our 14 subjects.

```
cor(breakfast$cornflakes, breakfast$oatbran)
```

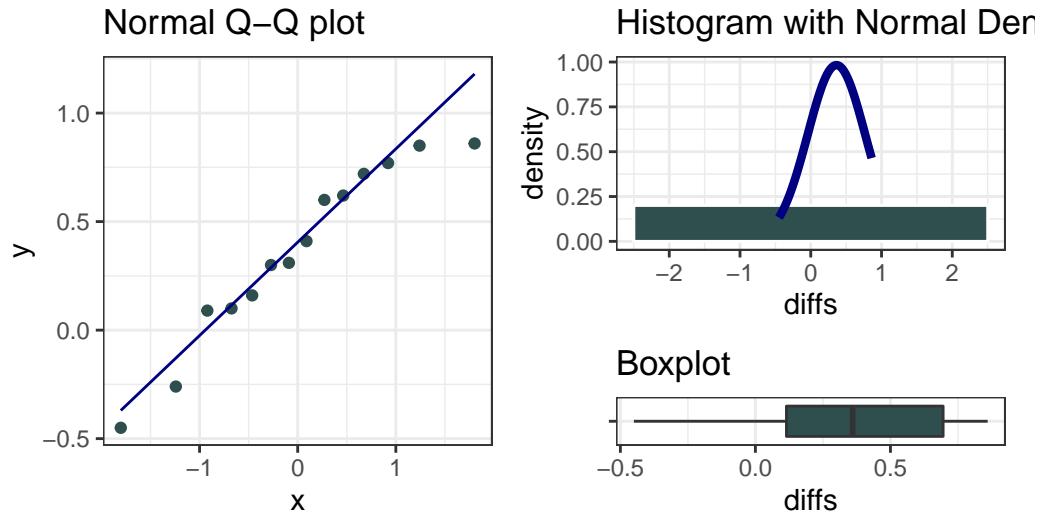
```
[1] 0.9233247
```

The sample Pearson correlation coefficient is very strong and positive at 0.92, so the paired samples approach will use these data far more effectively than the (incorrect) independent samples approach.

20.3.8 Question 8. What does the distribution of paired differences tell us?

We summarize the distribution of the paired differences (cornflakes - oatbran) below.

Difference in LDL (Corn Flakes – Oat Bran)



The Normal distribution doesn't look too ridiculous in this case for the paired (cornflakes-oatbran) differences. Suppose we assume Normality and run the paired t test.

```
t.test(breakfast$cornflakes - breakfast$oatbran)
```

One Sample t-test

```
data: breakfast$cornflakes - breakfast$oatbran
t = 3.3444, df = 13, p-value = 0.005278
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
0.1284606 0.5972537
sample estimates:
mean of x
0.3628571
```

Based on this sample of 14 subjects in the crossover study, we observe a 95% confidence interval for the difference between the LDL cholesterol levels after eating corn flakes and eating oat

bran that is entirely positive, suggesting that LDL levels were detectably higher (according to this t test procedure) after eating corn flakes than after eating oat bran.

20.4 Power, Sample Size and the Breakfast Study

As a preview of what's next to come, let's investigate these promising results a bit further. Suppose that in a new study, you wish to be able to detect a difference in LDL cholesterol between two exposures: subjects who eat cornflakes (as in the original study) and subjects who continue to eat cornflakes but also take a supplemental dosage of what you believe to be the crucial ingredient in oatbran.

Suppose you believe that the effect of taking the new supplement will be about half the size of the effect you observed in the original breakfast study on hypercholesterolemic males, but that males generally may be more likely to take your supplement regularly than switch from cornflakes to a less appetizing breakfast choice, making your supplement attractive.

What sample size will be required to yield 90% power to detect an effect half the size of the effect we observed in the breakfast study, in a new paired samples study using a two-tailed 5% significance level? What if we only required 80% power?

20.4.1 The Setup

We want to know n , the minimum required sample size for the new study, and we have:

- A specified effect size of half of what we saw in the breakfast study, where the sample mean difference between cornflakes and oatbran was 0.36 mmol/l, so our effect size is assumed to be `delta = 0.18 mmol/l`.
- An assumed standard deviation equal to the standard deviation of the differences in the pilot breakfast study, which turns out to have been $s = 0.41 \text{ mmol/l}$.
- We also have a pre-specified `alpha = 0.05` using a two-tailed test.
- We also want the power to be at least 90% for our new study.

20.4.2 The R Calculations

Question 1. What sample size will be required to yield 90% power to detect an effect half the size of the effect we observed in the breakfast study, in a new paired samples study using a two-tailed 5% significance level?

```
power.t.test(delta = 0.18, sd = 0.41, sig.level = 0.05,
              power = 0.9, type="paired", alternative="two.sided")
```

```
Paired t test power calculation
```

```
n = 56.47119
delta = 0.18
sd = 0.41
sig.level = 0.05
power = 0.9
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

And so our new study will require at least **57 subjects** (each measured in two circumstances, so 114 total measurements) in order to achieve at least 90% power to detect the difference of 0.18 mmol/l while meeting these specifications.

Question 2. What if we were willing to accept only 80% power?

```
power.t.test(delta = 0.18, sd = 0.41, sig.level = 0.05,
              power = 0.8, type="paired", alternative="two.sided")
```

```
Paired t test power calculation
```

```
n = 42.68269
delta = 0.18
sd = 0.41
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

It turns out that this would require at least **43 subjects**.

20.4.3 Independent samples, instead of paired samples?

What would happen if, instead of doing a paired samples study, we did one using independent samples? Assuming we used a balanced design, and assigned the same number of different people at random to either the oatbran supplement or regular cornflakes alone, we could do

such a study, but it would require many more people to obtain similar power to the paired samples study.

```
power.t.test(delta = 0.18, sd = 0.41, sig.level = 0.05,
              power = 0.9, type="two.sample", alternative="two.sided")
```

Two-sample t test power calculation

```
n = 110
delta = 0.18
sd = 0.41
sig.level = 0.05
power = 0.9
alternative = two.sided
```

NOTE: n is number in **each** group

In all, **220 people** would be required in the independent samples study (110 in each exposure group), as compared to only **57 people** (each measured twice) in the paired study.

21 Estimating a Population Proportion

We've focused on creating statistical inferences about a population mean, or difference between means, where we care about a quantitative outcome. Now, we'll tackle **categorical** outcomes, by estimating a confidence interval around a population proportion.

21.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

source("data/Love-boost.R")
library(janitor)
library(knitr)
library(broom)
library(tidyverse)

theme_set(theme_bw())
```

We will also use some functions from the `mosaic` package.

21.2 A First Example: Serum Zinc in the “Normal” Range?

Recall that in the serum zinc study, we have 462 teenage male subjects, of whom 395 (or 85.5%) fell in the “normal range” of 66 to 110 micrograms per deciliter.

```
serzinc <- read_csv("data/serzinc.csv",
                     show_col_types = FALSE)

serzinc <- serzinc |>
  mutate(in_range = ifelse(zinc >= 66 & zinc <= 110, 1, 0))

serzinc |> tabyl(in_range) |>
  adorn_totals() |> adorn_pct_formatting()
```

```

in_range   n percent
  0    67   14.5%
  1   395   85.5%
Total 462 100.0%

```

Previously, we estimated a confidence interval for the *mean* of the population zinc levels. Now, we want to estimate a confidence interval for the *proportion* of the population whose serum zinc levels are in the range of 66 to 110. We want to build both a point estimate for the population proportion, and a confidence interval for the population proportion.

Now, let's identify a 95% confidence interval for the proportion of the population whose zinc levels are within the “normal” range. We have seen that 395 / 462 subjects (or a proportion of 0.855) fall in the “normal range” in our sample. For now, that will also be our *point estimate* of the proportion in the “normal range” across the entire population of teenagers like those in our sample.

```

serzinc <- serzinc |>
  mutate(in_range = ifelse(zinc > 65 & zinc < 111, 1, 0))

serzinc |> tabyl(in_range)

in_range   n   percent
  0    67 0.1450216
  1   395 0.8549784

```

Once we've created this 0-1 variable, there are several available approaches for wrapping a confidence interval around this proportion.

21.2.1 Using an Intercept-Only Regression Again?

We might consider taking the same approach as we did with the population mean earlier:

```

model_zincprop <- lm(in_range ~ 1, data = serzinc)

tidy(model_zincprop, conf.int = TRUE, conf = 0.90) |>
  knitr::kable(digits = 3)

```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.855	0.016	52.133	0	0.828	0.882

While there are more powerful approaches to estimate a confidence interval around this proportion, this simple approach turns out not to be too bad, so long as the sample proportion isn't very close to either 0 or 1.

21.2.2 A $100(1-\alpha)\%$ Confidence Interval for a Population Proportion

Suppose we want to estimate a confidence interval for an unknown population proportion, π , on the basis of a random sample of n observations from that population which yields a sample proportion of p . Note that this p is the sample proportion – it's not a p value.

- In our serum zinc example, we have $n = 462$ observations, with a sample proportion ("in range") of $p = 0.855$.

A $100(1-\alpha)\%$ confidence interval for the population proportion π can be created by using the standard normal distribution, the sample proportion, p , and the standard error of a sample proportion, which is defined as the square root of p multiplied by $(1-p)$ divided by the sample size, n .

- So the standard error is estimated in our serum zinc example as:

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.855(1-0.855)}{462}} = \sqrt{0.000268} = 0.016$$

And thus, our confidence interval is $p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$

where $Z_{\alpha/2}$ = the value from a standard Normal distribution cutting off the top $\alpha/2$ of the distribution, obtained in R by substituting the desired $\alpha/2$ value into the following command: `qnorm(alpha/2, lower.tail=FALSE)`.

Note: This interval is reasonably accurate so long as np and $n(1-p)$ are each at least 5.

- For the serum zinc data, we have $np = (462)(0.855) = 395$ and $n(1-p) = 462(1 - 0.855) = 67$, so this should be ok.
- For $\alpha = 0.05$, we have $Z_{\alpha/2} = 1.96$, approximately.

```
qnorm(0.025, lower.tail = FALSE)
```

```
[1] 1.959964
```

- Thus, for the serum zinc estimate, this confidence interval would be:

$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} = \frac{395}{462} \pm 1.96 \sqrt{\frac{0.855(1-0.855)}{462}} = 0.855 \pm 0.032$$

or (0.823, 0.887).

21.3 Using `binom.test` from the `mosaic` package

I am aware of at least seven different procedures for estimating a confidence interval for a population proportion using R. All have minor weaknesses: none is importantly different from the others in many practical situations. Five of these methods are available using the `binom.test` function from the `mosaic` package in R.

The general format for using the `binom.test` function is as follows:

```
mosaic::binom.test(x = 395, # substitute in number of successes
                    n = 462, # substitute in number of trials
                    conf.level = 0.95, # default confidence level
                    p = 0.5, # default null hypothesis proportion
                    ci.method = "XXX") # see below for XXX options
```

where the appropriate `ci.method` is obtained from the table below.

Approach	<code>ci.method</code> to be used
Wald	“Wald”
Clopper-Pearson	“Clopper-Pearson” or “binom.test”
Score	“Score” or “prop.test”
Agresti-Coull	“agresti-coull”
Plus4	“plus4”

21.3.1 The Wald test approach

The Wald approach can be used to establish a very similar confidence interval to the one we calculated above, based on something called the Wald test.

Here, we specify the `x` and `n` values. `n` is the total number of observations, and `x` is the number where the event of interest (in this case, serum zinc levels in the normal range) occurs. So `x` = 395 and `n` = 462.

```
m_wald <- mosaic::binom.test(x = 395, n = 462,
                               conf.level = 0.95,
                               ci.method = "Wald")
```

```
Registered S3 method overwritten by 'mosaic':
  method                  from
  fortify.SpatialPolygonsDataFrame  ggplot2
```

```
m_wald
```

```
Exact binomial test (Wald CI)

data: 395 out of 462
number of successes = 395, number of trials = 462, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.8228698 0.8870869
sample estimates:
probability of success
 0.8549784
```

The Wald confidence interval is always symmetric around our point estimate, and can dip below 0 or above 1.

When I fit intervals using the approaches in `mosaic::binom_test()` I will usually tidy them.

```
tidy(m_wald) |>
  select(estimate, conf.low, conf.high, statistic, parameter) |>
  kable(digits = 3)
```

estimate	conf.low	conf.high	statistic	parameter
0.855	0.823	0.887	395	462

The other elements of the tidied result are shown below.

```
tidy(m_wald) |>
  select(method, alternative, p.value)
```

```
# A tibble: 1 x 3
  method           alternative   p.value
  <chr>            <chr>        <dbl>
1 Exact binomial test (Wald CI) two.sided 1.23e-57
```

21.3.2 The Clopper-Pearson approach

The `binom.test` command can be used to establish an “exact” confidence interval. This uses the method of Clopper and Pearson from 1934, and is exact in the sense that it guarantees, for instance, that the confidence level associated with the interval is at least as large as the nominal level of 95%, but not that the interval isn’t wider than perhaps it needs to be.

```
m_clopper <- mosaic::binom.test(x = 395, n = 462,
                                  conf.level = 0.95,
                                  ci.method = "Clopper-Pearson")
```

```
m_clopper
```

```
data: 395 out of 462
number of successes = 395, number of trials = 462, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.8195187 0.8858100
sample estimates:
probability of success
 0.8549784
```

Clopper-Pearson is used by `stats::binom.test()` in R as well. Again, it guarantees coverage at least as large as the nominal coverage rate, but may produce wider intervals than the other methods we’ll see. The 95% confidence interval by this method is (0.820, 0.886), which is in the same general range as our previous estimates.

```
tidy(m_clopper) |>
  select(estimate, conf.low, conf.high, statistic, parameter) |>
  kable(digits = 3)
```

estimate	conf.low	conf.high	statistic	parameter
0.855	0.82	0.886	395	462

21.3.3 The Score approach

```
m_score <- mosaic::binom.test(x = 395, n = 462,
                                conf.level = 0.95,
                                ci.method = "Score")
```

```
m_score
```

```
Exact binomial test (Score CI without continuity correction)

data: 395 out of 462
number of successes = 395, number of trials = 462, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.8199415 0.8841607
sample estimates:
probability of success
 0.8549784
```

The Score approach is also used by `stats::prop.test()` and creates CIs by inverting p-values from score tests. It can be applied with a continuity correction (use `ci.method = "prop.test"`) or without.

In this case, we see that the Score approach and the Clopper-Pearson approach give very similar results.

```
tidy(m_score) |>
  select(estimate, conf.low, conf.high, statistic, parameter) |>
  kable(digits = 3)
```

estimate	conf.low	conf.high	statistic	parameter
0.855	0.82	0.884	395	462

As mentioned, the score test can also be run incorporating something called a *continuity correction*, since we are using a Normal approximation to the exact binomial distribution to

establish our margin for error. R, by default, includes this continuity correction for the Score test when we use `prop.test` to collect it.

```
m_score_cor <- mosaic::binom.test(x = 395, n = 462,
                                    conf.level = 0.95,
                                    ci.method = "prop.test")
```

```
m_score_cor
```

```
Exact binomial test (Score CI with continuity correction)

data: 395 out of 462
number of successes = 395, number of trials = 462, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.8187706 0.8851359
sample estimates:
probability of success
0.8549784
```

```
tidy(m_score_cor) |>
  select(estimate, conf.low, conf.high, statistic, parameter) |>
  kable(digits = 3)
```

estimate	conf.low	conf.high	statistic	parameter
0.855	0.819	0.885	395	462

21.3.4 The Agresti-Coull Approach

```
m_agresti <- mosaic::binom.test(x = 395, n = 462,
                                   conf.level = 0.95,
                                   ci.method = "agresti-coull")
```

```
m_agresti
```

```
Exact binomial test (Agresti-Coull CI)
```

```

data: 395 out of 462
number of successes = 395, number of trials = 462, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.8198094 0.8842928
sample estimates:
probability of success
 0.8549784

```

The Agresti-Coull approach is the Wald method after adding Z successes and Z failures to the data, where Z is the appropriate quantile for a standard Normal distribution (1.96 for a 95% CI).

```

tidy(m_agresti) |>
  select(estimate, conf.low, conf.high, statistic, parameter) |>
  kable(digits = 3)

```

estimate	conf.low	conf.high	statistic	parameter
0.855	0.82	0.884	395	462

21.3.5 The “Plus 4” approach

This approach is just the Wald method after adding 2 successes and 2 failures (so 4 observations) to the data. It will be very similar to the Agresti-Coull method if we are working with a 95% confidence interval.

```

m_plus4 <- mosaic::binom.test(x = 395, n = 462,
                                conf.level = 0.95,
                                ci.method = "plus4")

m_plus4

```

```

Exact binomial test (Plus 4 CI)

data: 395 out of 462
number of successes = 395, number of trials = 462, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5

```

```

95 percent confidence interval:
 0.8196844 0.8841783
sample estimates:
probability of success
 0.8549784

```

```

tidy(m_plus4) |>
  select(estimate, conf.low, conf.high, statistic, parameter) |>
  kable(digits = 3)

```

estimate	conf.low	conf.high	statistic	parameter
0.855	0.82	0.884	395	462

21.3.6 SAIFS: single augmentation with an imaginary failure or success

SAIFS stands for “single augmentation with an imaginary failure or success” and the method I’ll describe is one of several similar approaches. The next subsection describes the R code for calculating the relevant confidence interval.

An approach I like for the estimation of a confidence interval for a single population proportion/rate¹ is to estimate the lower bound of a confidence interval with an imaginary failure added to the observed data, and estimate the upper bound of a confidence interval with an imaginary success added to the data.

Suppose we have X successes in n trials, and we want to establish a confidence interval for the population proportion of successes.

Let $p_1 = (X + 0)/(n + 1)$, $p_2 = (X + 1)/(n + 1)$, $q_1 = 1 - p_1$, $q_2 = 1 - p_2$

- The lower bound of a $100(1-\alpha)\%$ confidence interval for the population proportion of successes using the SAIFS procedure is then $LB_{SAIFS}(x, n, \alpha) = p_1 - t_{\alpha/2, n-1} \sqrt{\frac{p_1 q_1}{n}}$
- The upper bound of that same $100(1-\alpha)\%$ confidence interval for the population proportion of successes using the SAIFS procedure is $UB_{SAIFS}(x, n, \alpha) = p_2 + t_{\alpha/2, n-1} \sqrt{\frac{p_2 q_2}{n}}$

Returning to the serum zinc example, we’ve got 395 “successes” (value in the normal range) out of 462 “trials” (values measured), so that $X = 395$ and $n = 462$

¹See Borkowf CB (2006) Constructing binomial confidence intervals with near nominal coverage by adding a single imaginary failure or success. Statistics in Medicine. 25(21): 3679-3695. doi: 10.1002/sim.2469, or get the whole PDF of the paper at <http://onlinelibrary.wiley.com/doi/10.1002/sim.2469/pdf>

So we have $p_1 = \frac{X+0}{n+1} = \frac{395}{463} = 0.8531$, $p_2 = \frac{X+1}{n+1} = \frac{396}{463} = 0.8553$, and $q_1 = 1 - p_1 = 0.1469$ and $q_2 = 1 - p_2 = 0.1447$

We have $n = 462$ so if we want a 95% confidence interval ($\alpha = 0.05$), then we have $t_{\alpha/2, n-1} = t_{0.025, 461} = 1.9651$, which I determined using R's `qt` function:

```
qt(0.025, df = 461, lower.tail=FALSE)
```

```
[1] 1.965123
```

- Thus, our lower bound for a 95% confidence interval is $p_1 - t_{\alpha/2, n-1} \sqrt{\frac{p_1 q_1}{n}}$, or $0.8531 - 1.9651 \sqrt{\frac{0.8531(0.1469)}{462}}$, which is $0.8531 - 0.0324$ or 0.8207 .
- Our upper bound is $p_2 + t_{\alpha/2, n-1} \sqrt{\frac{p_2 q_2}{n}}$, or $0.8553 + 1.9651 \sqrt{\frac{0.8553(0.1447)}{462}}$, which is $0.8553 + 0.0323$, or 0.8876 .

So the 95% SAIFS confidence interval estimate for the population proportion, π , of teenage males whose serum zinc levels fall within the “normal range” is $(0.821, 0.888)$.

21.3.7 A Function in R to Calculate the SAIFS Confidence Interval

I built an R function, called `saifs.ci` and contained in the Markdown for this document as well as the `Love-boost.R` script on the web site, which takes as its arguments a value for X = the number of successes, n = the number of trials, and `conf.level` = the confidence level, and produces the sample proportion, the SAIFS lower bound and upper bound for the specified two-sided confidence interval for the population proportion, using the equations above.

Here, for instance, are 95%, 90% and 99% confidence intervals for the population proportion π that we have been studying in the serum zinc data.

```
saifs.ci(x = 395, n = 462)
```

Sample Proportion	0.025	0.975
0.855	0.821	0.887

```
saifs.ci(x = 395, n = 462, conf=0.9)
```

Sample Proportion	0.05	0.95
0.855	0.826	0.882

```
saifs.ci(x = 395, n = 462, conf=0.99, dig=5)
```

Sample Proportion	0.005	0.995
0.85498	0.81054	0.89763

Note that in the final interval, I asked the machine to round to five digits rather than the default of three. On my desktop (and probably yours), doing so results in this output:

Sample Proportion	0.005	0.995
0.85498	0.81054	0.89763

I've got some setting wrong in my bookdown work so that this doesn't show up above when the function is called. Sorry!

21.3.8 The `saifs.ci` function in R

```
`saifs.ci` <-  
  function(x, n, conf.level=0.95, dig=3)  
  {  
    p.sample <- round(x/n, digits=dig)  
  
    p1 <- x / (n+1)  
    p2 <- (x+1) / (n+1)  
  
    var1 <- (p1*(1-p1))/n  
    se1 <- sqrt(var1)  
    var2 <- (p2*(1-p2))/n  
    se2 <- sqrt(var2)  
  
    lowq = (1 - conf.level)/2  
    tcut <- qt(lowq, df=n-1, lower.tail=FALSE)  
  
    lower.bound <- round(p1 - tcut*se1, digits=dig)  
    upper.bound <- round(p2 + tcut*se2, digits=dig)  
    res <- c(p.sample, lower.bound, upper.bound)  
    names(res) <- c('Sample Proportion',lowq, 1-lowq)  
    res  
  }
```

21.4 A Second Example: Ebola Mortality Rates through 9 Months of the Epidemic

The World Health Organization's Ebola Response Team published an article² in the October 16, 2014 issue of the New England Journal of Medicine, which contained some data I will use in this example, focusing on materials from their Table 2.

As of September 14, 2014, a total of 4,507 confirmed and probable cases of Ebola virus disease (EVD) had been reported from West Africa. In our example, we will look at a set of 1,737 cases, with definitive outcomes, reported in Guinea, Liberia, Nigeria and Sierra Leone.

Across these 1,737 cases, a total of 1,229 cases led to death. Based on these sample data, what can be said about the case fatality rate in the population of EVD cases with definitive outcomes for this epidemic?

21.4.1 Working through the Ebola Virus Disease Example

We have $n = 1,737$ subjects, of whom we observed death in 1,229, for a sample proportion of $p = \frac{1229}{1737} = 0.708$. The standard error of that sample proportion will be

$$SE(p) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.708(1-0.708)}{1737}} = 0.011$$

And our 95% confidence interval (so that we'll use $\alpha = 0.05$) for the true population proportion, π , of EVD cases with definitive outcomes, who will die is $p \pm Z_{0.025} \sqrt{\frac{p(1-p)}{n}}$, or $0.708 \pm 1.96(0.011) = 0.708 \pm 0.022$, or (0.686, 0.730)

Note that I simply recalled from our prior work that $Z_{0.025} = 1.96$, but we can verify this:

```
qnorm(0.025, lower.tail=FALSE)
```

```
[1] 1.959964
```

Since both $np=(1737)(0.708)=1230$ and $n(1-p)=(1737)(1-0.708)=507$ are substantially greater than 5, this should be a reasonably accurate confidence interval.

We have 95% confidence in an interval estimate for the true population proportion that falls between 0.686 and 0.730. Equivalently, we could say that we're 95% confident that the true case fatality rate expressed as a percentage rather than a proportion, is between 68.6% and 73.0%.

²WHO Ebola Response Team (2014) Ebola virus disease in West Africa: The first 9 months of the epidemic and forward projections. New Engl J Med 371: 1481-1495 doi: 10.1056/NEJMoa1411100

21.4.2 Using R to estimate the CI for our Ebola example

```
ebola_wald <- mosaic::binom.test(x = 1229, n = 1737, conf.level = 0.95,
                                    ci.method = "Wald") |>
  tidy() |> select(estimate, conf.low, conf.high)
ebola_clop <- mosaic::binom.test(x = 1229, n = 1737, conf.level = 0.95,
                                    ci.method = "Clopper-Pearson") |>
  tidy() |> select(estimate, conf.low, conf.high)
ebola_scor <- mosaic::binom.test(x = 1229, n = 1737, conf.level = 0.95,
                                    ci.method = "Score") |>
  tidy() |> select(estimate, conf.low, conf.high)
ebola_agco <- mosaic::binom.test(x = 1229, n = 1737, conf.level = 0.95,
                                    ci.method = "agresti-coull") |>
  tidy() |> select(estimate, conf.low, conf.high)
ebola_plus <- mosaic::binom.test(x = 1229, n = 1737, conf.level = 0.95,
                                    ci.method = "plus4") |>
  tidy() |> select(estimate, conf.low, conf.high)

ebola_res <- bind_rows(ebola_wald, ebola_clop, ebola_scor,
                       ebola_agco, ebola_plus) |>
  mutate(approach = c("Wald", "Clopper-Pearson", "Score",
                      "Agresti-Coull", "Plus4"))

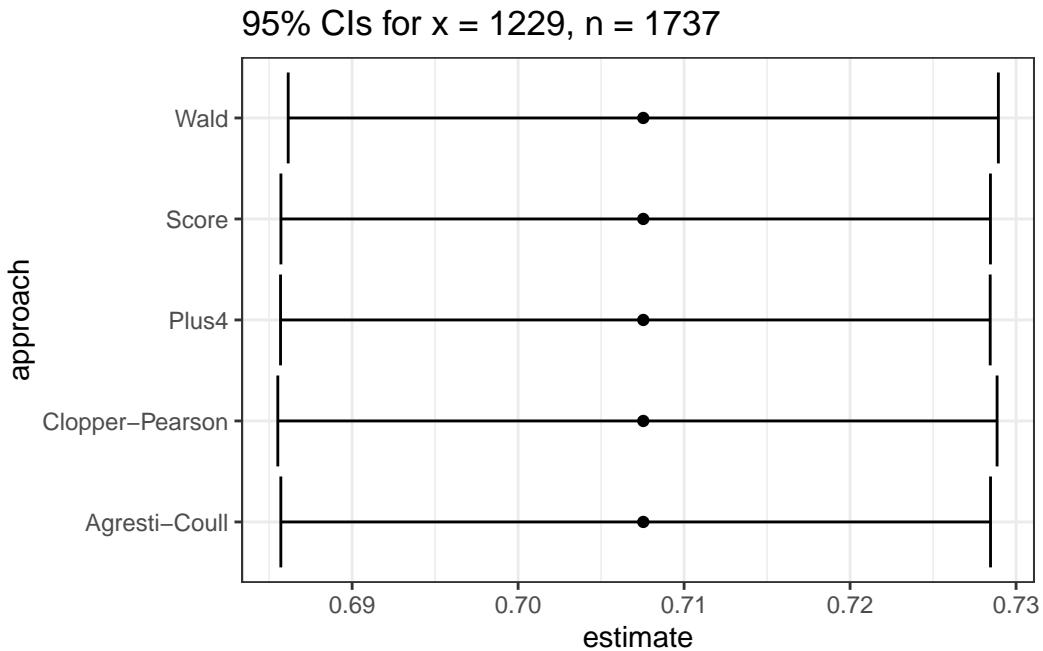
ebola_res |> kable(digits = 6)
```

estimate	conf.low	conf.high	approach
0.707542	0.686150	0.728934	Wald
0.707542	0.685524	0.728856	Clopper-Pearson
0.707542	0.685710	0.728457	Score
0.707542	0.685705	0.728462	Agresti-Coull
0.707542	0.685687	0.728443	Plus4

This is way more precision than we can really justify, but I just want you to see that the five results are all (slightly) different.

21.4.3 Plotting the Confidence Intervals for the Ebola Virus Disease Example

```
ggplot(ebola_res, aes(x = approach, y = estimate)) +  
  geom_point() +  
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high)) +  
  coord_flip() +  
  labs(title = "95% CIs for x = 1229, n = 1737")
```



So in this case, it really doesn't matter which one you choose. With a smaller sample, we may not come to the same conclusion about the relative merits of these different approaches.

21.4.4 What about the `saifs.ci()` result?

```
saifs.ci(x = 1229, n = 1737, conf.level=0.95)
```

Sample Proportion	0.025	0.975
	0.708	0.729
0.686		

21.5 Can the Choice of Confidence Interval Method Matter?

Yes. This will especially be the case when we have a small sample size, and a probability of “success” that is close to either 0 or 1. For instance, suppose we run 10 trials, and obtain a single success, then use these data to estimate the true proportion of success, π .

The 90% confidence intervals under this circumstance are very different.

```
tidy1 <- mosaic::binom.test(x = 1, n = 10, conf.level = 0.90,
                             ci.method = "Wald") |> tidy()
tidy2 <- mosaic::binom.test(x = 1, n = 10, conf.level = 0.90,
                             ci.method = "Clopper-Pearson") |> tidy()
tidy3 <- mosaic::binom.test(x = 1, n = 10, conf.level = 0.90,
                             ci.method = "Score") |> tidy()
tidy4 <- mosaic::binom.test(x = 1, n = 10, conf.level = 0.90,
                             ci.method = "agresti-coull") |> tidy()
tidy5 <- mosaic::binom.test(x = 1, n = 10, conf.level = 0.90,
                             ci.method = "plus4") |> tidy()

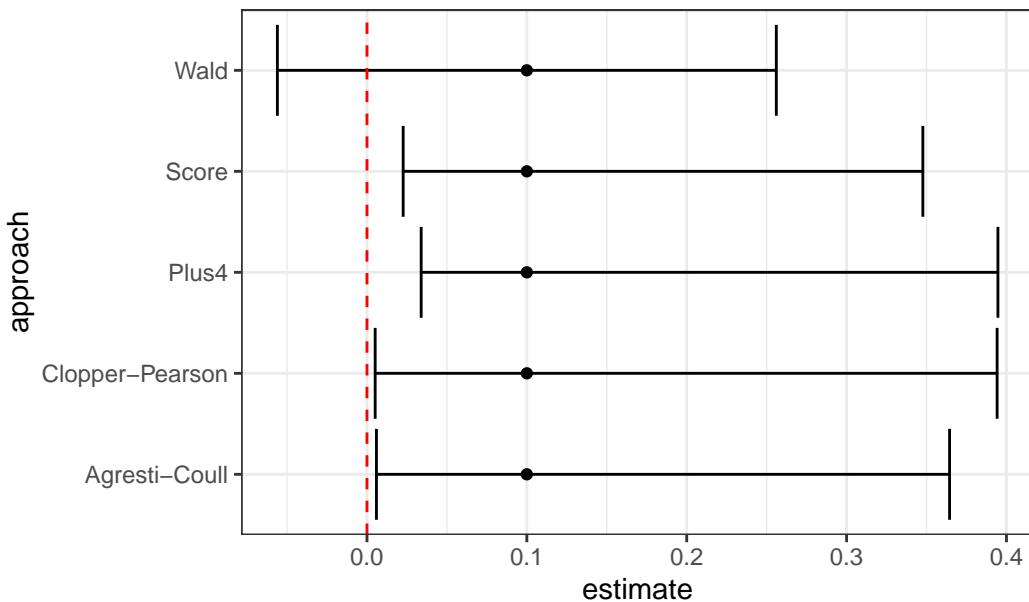
res <- bind_rows(tidy1, tidy2, tidy3, tidy4, tidy5) |>
  mutate(approach = c("Wald", "Clopper-Pearson", "Score",
                     "Agresti-Coull", "Plus4")) |>
  select(approach, estimate, conf.low, conf.high)

res |> kable(digits = 3)
```

approach	estimate	conf.low	conf.high
Wald	0.1	-0.056	0.256
Clopper-Pearson	0.1	0.005	0.394
Score	0.1	0.023	0.348
Agresti-Coull	0.1	0.006	0.364
Plus4	0.1	0.034	0.395

Note that the Wald procedure doesn’t force the confidence interval to appear in the (0, 1) range.

90% CIs for $x = 1$, $n = 10$



None of these three approaches is always better than any of the others. When we have a sample size below 100, or the sample proportion of success is either below 0.10 or above 0.90, caution is warranted, although in many cases, the various methods give similar responses.

22 Comparing Proportions with Two Independent Samples

Often, when analyzing data from an independent samples design, we are interested in comparing the proportions of subjects that achieve some outcome, across two levels of an exposure, or treatment. In this circumstance, we will first summarize the available data in terms of a 2x2 cross-tabulation, and then apply a series of inferential methods for 2x2 tables to obtain point and interval estimates of interest.

22.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

source("data/Love-boost.R")
library(janitor)
library(tidyverse)

theme_set(theme_bw())
```

We will also use some functions from the `mosaic` package.

22.2 A First Example: Ibuprofen and Sepsis Trial

As we saw in a previous Chapter, we are interested in comparing the percentage of patients in each arm of the trial (Ibuprofen vs. Placebo) that showed an improvement in their temperature (`temp_drop > 0`). Our primary interest is in comparing the percentage of Ibuprofen patients whose temperature dropped to the percentage of Placebo patients whose temperature dropped.

- We can summarize the data behind the two proportions we are comparing in a contingency table with two rows which identify the exposure or treatment of interest, and two columns to represent the outcomes of interest.

- In this case, we are comparing two groups of subjects based on treatment (`treat`): those who received Ibuprofen and those who received a placebo. The outcome of interest is whether the subject's temperature dropped (`temp_drop > 0`), or not.
- In the table, we place the frequency for each combination of a row and a column.
- The rows need to be mutually exclusive and collectively exhaustive: each patient must either receive Ibuprofen or Placebo. Similarly, the columns must meet the same standard: every patient's temperature either drops or does not drop.

Here's the contingency table.

```

sepsis <- read_csv("data/sepsis.csv",
                     show_col_types = FALSE) |>
  mutate(treat = factor(treat),
         race = factor(race))

sepsis <- sepsis |>
  mutate(dropped = ifelse(temp_drop > 0, "Drop", "No Drop"))

sepsis |> tabyl(treat, dropped) |>
  adorn_totals() |>
  adorn_percentages(denom = "row") |>
  adorn_pct_formatting(digits = 1) |>
  adorn_ns(position = "front")

      treat      Drop     No Drop
Ibuprofen 107 (71.3%) 43 (28.7%)
Placebo   80 (53.3%) 70 (46.7%)
Total    187 (62.3%) 113 (37.7%)
  
```

In our sample, we observe 71.3% of the 150 Ibuprofen subjects with a positive `temp_drop` as compared to 53.3% of the 150 Placebo subjects. We want to compare these two probabilities (represented using proportions as 0.713 vs. 0.533) to estimate the size of the difference between the proportions with a point estimate and 90% confidence interval.

But there are other comparisons we could make, too. The tricky part is that we have multiple ways to describe the relationship between treatment and outcome. We might compare outcome “risks” directly using the difference in probabilities, or the ratio of the two probabilities, or we might convert the risks to odds, and compare the ratio of those odds. In any case, we'll get different point estimates and confidence intervals, all of which will help us make conclusions about the evidence available in this trial speaking to differences between Ibuprofen and Placebo.

22.3 Relating a Treatment to an Outcome

The question of interest is whether the percentage of subjects whose temperature dropped is different (and probably larger) in the subjects who received Ibuprofen than in those who received the Placebo.

Treatment Arm	Did Not			Proportion who dropped
	Dropped	Drop	Total	
Ibuprofen	107	43	150	0.713
Placebo	80	70	150	0.533

In other words, what is the relationship between the treatment and the outcome?

22.4 Definitions of Probability and Odds

- Proportion = Probability = Risk of the trait = number with trait / total
- Odds of having the trait = (number with the trait / number without the trait) to 1

If p is the proportion of subjects with a trait, then the **odds** of having the trait are $\frac{p}{1-p}$ to 1.

So, the probability of a good result (temperature drop) in this case is $\frac{107}{150} = 0.713$ in the Ibuprofen group. The **odds** of a good result are thus $\frac{0.713}{1-0.713} = 2.484$ to 1 in the Ibuprofen group.

Treatment Arm	Did Not			Pr(dropped)	Odds(dropped)
	Dropped	Drop	Total		
Ibuprofen	107	43	150	0.713	2.484
Placebo	80	70	150	0.533	1.141

22.5 Defining the Relative Risk

Among the Ibuprofen subjects, the risk of a good outcome (drop in temperature) is 71.3% or, stated as a proportion, 0.713. Among the Placebo subjects, the risk of a good outcome is 53.3% or, stated as a proportion, 0.533.

Our “crude” estimate of the **relative risk** of a good outcome for Ibuprofen subjects as compared to Placebo subjects, is the ratio of these two risks, or $0.713/0.533 = 1.338$

- The fact that this relative risk is greater than 1 indicates that the probability of a good outcome is higher for Ibuprofen subjects than for Placebo subjects.
- A relative risk of 1 would indicate that the probability of a good outcome is the same for Ibuprofen subjects and for Placebo subjects.
- A relative risk less than 1 would indicate that the probability of a good outcome is lower for Ibuprofen subjects than for Placebo subjects.

22.6 Defining the Risk Difference

Our “crude” estimate of the **risk difference** of a good outcome for Ibuprofen subjects as compared to Placebo subjects, is $0.713 - 0.533 = 0.180$ or 18.0 percentage points.

- The fact that this risk difference is greater than 0 indicates that the probability of a good outcome is higher for Ibuprofen subjects than for Placebo subjects.
- A risk difference of 0 would indicate that the probability of a good outcome is the same for Ibuprofen subjects and for Placebo subjects.
- A risk difference less than 0 would indicate that the probability of a good outcome is lower for Ibuprofen subjects than for Placebo subjects.

22.7 Defining the Odds Ratio, or the Cross-Product Ratio

Among the Ibuprofen subjects, the odds of a good outcome (temperature drop) are 2.484. Among the placebo subjects, the odds of a good outcome (temperature drop) are 1.141.

So our “crude” estimate of the **odds ratio** of a good outcome for Ibuprofen subjects as compared to placebo subjects, is $2.484 / 1.141 = 2.18$

Another way to calculate this odds ratio is to calculate the **cross-product ratio**, which is equal to $(a \times d) / (b \times c)$, for the 2 by 2 table with counts specified as shown:

A Generic Table	Good Outcome	Bad Outcome
Treatment Group 1	a	b
Treatment Group 2	c	d

So, for our table, we have $a = 107$, $b = 43$, $c = 80$, and $d = 70$, so the cross-product ratio is $\frac{107 \times 70}{43 \times 80} = \frac{7490}{3440} = 2.18$. As expected, this is the same as the “crude” odds ratio estimate.

- The fact that this odds ratio risk is greater than 1 indicates that the odds of a good outcome are higher for Ibuprofen subjects than for Placebo subjects.

- An odds ratio of 1 would indicate that the odds of a good outcome are the same for Ibuprofen subjects and for Placebo subjects.
- An odds ratio less than 1 would indicate that the odds of a good outcome are lower for Ibuprofen subjects than for Placebo subjects.

So, we have several different ways to compare the outcomes across the treatments. Are these differences and ratios large enough to rule out chance?

22.8 Comparing Rates in a 2x2 Table

What is the relationship between the treatment (Ibuprofen vs. Placebo) and the outcome (drop in temperature) in the following two-by-two table?

22.9 The twobytwo function in R

I built the `twobytwo` function in R (based on existing functions in the `Epi` library, which you need to have in your installed packages list in order for this to work) to do the work for this problem. All that is required is a single command, and a two-by-two table in standard epidemiological format (with the outcomes in the columns, and the treatments in the rows.)

Treatment Arm	Dropped	Did Not Drop
Ibuprofen	107	43
Placebo	80	70

The command just requires you to read off the cells of the table, followed by the labels for the two treatments, then the two outcomes, then a specification of the names of the rows (exposures) and columns (outcomes) from the table, and a specification of the confidence level you desire. We'll use 90% here.

The resulting output follows.

```
twobytwo(107, 43, 80, 70,
         "Ibuprofen", "Placebo", "Dropped", "No Drop",
         conf.level = 0.90)
```

2 by 2 table analysis:

```
-----  
Outcome : Dropped  
Comparing : Ibuprofen vs. Placebo
```

	Dropped	No Drop	P(Dropped)	90% conf.	interval
Ibuprofen	107	43	0.7133	0.6490	0.7701
Placebo	80	70	0.5333	0.4661	0.5993

	90% conf. interval	
Relative Risk:	1.3375	1.1492 1.5567
Sample Odds Ratio:	2.1773	1.4583 3.2509
Conditional MLE Odds Ratio:	2.1716	1.4177 3.3437
Probability difference:	0.1800	0.0881 0.2677

Exact P-value:	0.0019
Asymptotic P-value:	0.0014

22.9.1 Standard Epidemiological Format

This table is in **standard epidemiological format**, which means that:

- The rows of the table describe the “treatment” (which we’ll take here to be `treat`). The more interesting (sometimes also the more common) “treatment” is placed in the top row. That’s Ibuprofen here.
- The columns of the table describe the “outcome” (which we’ll take here to be whether the subject’s temperature dropped.) Typically, the more common “outcome” is placed to the left.

22.9.2 Outcome Probabilities and Confidence Intervals Within the Treatment Groups

The `twobytwo` output starts with estimates of the probability (risk) of a “Dropped” outcome among subjects who fall into the two treatment groups (Ibuprofen or Placebo), along with 90% confidence intervals for each of these probabilities.

2 by 2 table analysis:

Outcome : Dropped
Comparing : Ibuprofen vs. Placebo

	Dropped	No Drop	P(Dropped)	90% conf.	interval
Ibuprofen	107	43	0.7133	0.6490	0.7701
Placebo	80	70	0.5333	0.4661	0.5993

The conditional probability of a temperature drop given that the subject is in the Ibuprofen group, is symbolized as $\Pr(\text{Dropped} \mid \text{Ibuprofen}) = 0.7133$, and the 90% confidence interval around that proportion is $(0.6490, 0.7701)$.

- Note that these two confidence intervals fail to overlap, and so we expect to see a fairly large difference in the estimated probability of a temperature drop when we compare Ibuprofen to Placebo.

22.9.3 Relative Risk, Odds Ratio and Risk Difference, with Confidence Intervals

These elements are followed by estimates of the relative risk, odds ratio, and risk difference, each with associated 90% confidence intervals.

	90% conf. interval	
Relative Risk:	1.3375	1.1492 1.5567
Sample Odds Ratio:	2.1773	1.4583 3.2509
Conditional MLE Odds Ratio:	2.1716	1.4177 3.3437
Probability difference:	0.1800	0.0881 0.2677

- The **relative risk**, or the ratio of $P(\text{Temperature Drop} \mid \text{Ibuprofen})$ to $P(\text{Temperature Drop} \mid \text{Placebo})$, is shown first. Note that the 90% confidence interval is entirely greater than 1.
- The **odds ratio** is presented using two different definitions (the sample odds ratio is the cross-product ratio we mentioned earlier). Note that the 90% confidence interval using either approach is entirely greater than 1.
- The **probability (or risk) difference** $[P(\text{Temperature Drop} \mid \text{Ibuprofen}) - P(\text{Temperature Drop} \mid \text{Placebo})]$ is presented last. Note that the 90% confidence interval is entirely greater than 0.
- Note carefully that if there had been no difference between Ibuprofen and Placebo, the relative risk and odds ratios would be 1, but the probability difference would be zero.

22.10 Estimating a Rate More Accurately: Use $(x + 2)/(n + 4)$ rather than x/n

Suppose you have some data involving n independent tries, with x successes. A natural estimate of the “success rate” in the data is x / n .

But, strangely enough, it turns out this isn’t an entirely satisfying estimator. Alan Agresti provides substantial motivation for the $(x + 2)/(n + 4)$ estimate as an alternative¹. This is sometimes called a *Bayesian augmentation*.

¹This note comes largely from a May 15 2007 entry in Andrew Gelman’s blog at <http://andrewgelman.com/2007/05/15>

- The big problem with x / n is that it estimates $p = 0$ or $p = 1$ when $x = 0$ or $x = n$.
- It's also tricky to compute confidence intervals at these extremes, since the usual standard error for a proportion, $\sqrt{np(1-p)}$, gives zero, which isn't quite right.
- $(x + 2)/(n + 4)$ is much cleaner, especially when you build a confidence interval for the rate.
- The only place where $(x + 2)/(n + 4)$ will go wrong is if n is small and the true probability is very close to 0 or 1.

For example, if $n = 10$, and p is 1 in a million, then x will almost certainly be zero, and an estimate of $1/12$ is much worse than the simple $0/10$. However, how big a deal is this? If p might be 1 in a million, you're not going to estimate it with a $n = 10$ experiment².

Applying this method to our Ibuprofen and Sepsis Trial data, we would simply add two to each frequency in the main four cells in our 2x2 table.

So instead of using

```
twobytwo(107, 43, 80, 70,
         "Ibuprofen", "Placebo", "Dropped", "No Drop",
         conf.level = 0.90)
```

the Bayesian augmentation would encourage us to look at

```
twobytwo(109, 45, 82, 72,
         "Ibuprofen", "Placebo", "Dropped", "No Drop",
         conf.level = 0.90)
```

2 by 2 table analysis:

Outcome : Dropped
Comparing : Ibuprofen vs. Placebo

	Dropped	No Drop	P(Dropped)	90% conf. interval
Ibuprofen	109	45	0.7078	0.6441 0.7643
Placebo	82	72	0.5325	0.4662 0.5977
<hr/>				
			90% conf. interval	
Relative Risk:	1.3293		1.1434 1.5453	
Sample Odds Ratio:	2.1268		1.4337 3.1550	
Conditional MLE Odds Ratio:	2.1215		1.3950 3.2421	
Probability difference:	0.1753		0.0845 0.2622	

²Andrew Gelman's example is "I'm not going to try ten 100-foot golf putts, miss all of them, and then estimate my probability of success as $1/12$."

Exact P-value: 0.0022
Asymptotic P-value: 0.0016

As you can see, the odds ratio and relative risk estimates are (a little) closer to 1, and the probability difference is also a little closer to 0. The Bayesian augmentation provides a slightly more conservative set of estimates of the impact of Ibuprofen as compared to Placebo.

It is likely that the augmented version is a more accurate estimate here, but the two estimates will be comparable, generally, so long as either (a) the sample size in each exposure group is more than, say, 30 subjects, and/or (b) the sample probability of the outcome is between 10% and 90% in each exposure group.

22.11 A Second Example: Ebola Virus Disease Study, again

For instance, recall the Ebola Virus Disease study from the *New England Journal of Medicine* that we described in the previous Chapter. Suppose we want to compare the proportion of deaths among cases that had a definitive outcome who were hospitalized to the proportion of deaths among cases that had a definitive outcome who were not hospitalized.

The article suggests that of the 1,737 cases with a definitive outcome, there were 1,153 hospitalized cases. Across those 1,153 hospitalized cases, 741 people (64.3%) died, which means that across the remaining 584 non-hospitalized cases, 488 people (83.6%) died.

Here is the initial contingency table, using only the numbers from the previous paragraph.

Initial Ebola Table	Deceased	Alive	Total
Hospitalized	741	—	1153
Not Hospitalized	488	—	584
Total			1737

Now, we can use arithmetic to complete the table, since the rows and the columns are each mutually exclusive and collectively exhaustive.

Ebola 2x2 Table	Deceased	Alive	Total
Hospitalized	741	412	1153
Not Hospitalized	488	96	584
Total	1229	508	1737

We want to compare the fatality risk (probability of being in the deceased column) for the population of people in the hospitalized row to the population of people in the not hospitalized row.

We can run these data through R, using the Bayesian augmentation (adding a death and a survival to the hospitalized and also to the not hospitalized groups.) We'll use a 95% confidence level this time.

```
twobytwo(741+2, 412+2, 488+2, 96+2,  
         "Hospitalized", "Not Hospitalized", "Deceased", "Alive",  
         conf.level = 0.95)
```

2 by 2 table analysis:

Outcome : Deceased
Comparing : Hospitalized vs. Not Hospitalized

	Deceased	Alive	P(Deceased)	95% conf. interval
Hospitalized	743	414	0.6422	0.6141 0.6693
Not Hospitalized	490	98	0.8333	0.8010 0.8613

	95% conf. interval		
Relative Risk:	0.7706	0.7285	0.8151
Sample Odds Ratio:	0.3589	0.2801	0.4599
Conditional MLE Odds Ratio:	0.3591	0.2772	0.4624
Probability difference:	-0.1912	-0.2307	-0.1490

Exact P-value:	0.0000
Asymptotic P-value:	0.0000

I'll leave it as an exercise for you to interpret these results and draw some conclusions.

23 Power and Proportions

23.1 Setup: Packages Used Here

```
source("data/Love-boost.R")
```

Again, we'll use this for the `twobytwo` function. We'll also use a couple of functions from the `pwr` package.

23.2 Tuberculosis Prevalence Among IV Drug Users

Consider a study to investigate factors affecting tuberculosis prevalence among intravenous drug users. The original data source is Graham NMH et al. (1992) Prevalence of Tuberculin Positivity and Skin Test Anergy in HIV-1-Seropositive and Seronegative Intravenous Drug Users. *JAMA*, 267, 369-373. Among 97 individuals who admit to sharing needles, 24 (24.7%) had a positive tuberculin skin test result; among 161 drug users who deny sharing needles, 28 (17.4%) had a positive test result. To start, we'll test the null hypothesis that the proportions of intravenous drug users who have a positive tuberculin skin test result are identical for those who share needles and those who do not.

```
twobytwo(24, 73, 28, 133,
         "Sharing Needles", "Not Sharing",
         "TB test+", "TB test-")
```

2 by 2 table analysis:

```
-----  
Outcome : TB test+  
Comparing : Sharing Needles vs. Not Sharing
```

	TB test+	TB test-	P(TB test+)	95% conf. interval
Sharing Needles	24	73	0.2474	0.1717 0.3427
Not Sharing	28	133	0.1739	0.1229 0.2404

95% conf. interval

```

Relative Risk: 1.4227      0.8772      2.3073
Sample Odds Ratio: 1.5616    0.8439      2.8898
Conditional MLE Odds Ratio: 1.5588    0.8014      3.0191
Probability difference: 0.0735   -0.0265     0.1807

Exact P-value: 0.1996
Asymptotic P-value: 0.1557
-----
```

What conclusions should we draw?

23.3 Designing a New TB Study

Now, suppose we wanted to design a new study with as many non-sharers as needle-sharers participating, and suppose that we wanted to detect any difference in the proportion of positive skin test results between the two groups that was identical to the data presented above or larger with at least 90% power, using a two-sided test and $\alpha = .05$. What sample size would be required to accomplish these aims?

23.4 Using `power.prop.test` for Balanced Designs

Our constraints are that we want to find the sample size for a two-sample comparison of proportions using a balanced design, we will use $\alpha = .05$, and power = .90, and that we estimate that the non-sharers will have a .174 proportion of positive tests, and we will try to detect a difference between this group and the needle sharers, who we estimate will have a proportion of .247, using a two-sided hypothesis test.

```
power.prop.test(p1 = .174, p2 = .247, sig.level = 0.05, power = 0.90)
```

```
Two-sample comparison of proportions power calculation
```

```

n = 653.2876
p1 = 0.174
p2 = 0.247
sig.level = 0.05
power = 0.9
alternative = two.sided
```

NOTE: n is number in *each* group

So, we'd need at least 654 non-sharing subjects, and 654 more who share needles to accomplish the aims of the study.

23.5 How power.prop.test works

power.prop.test works much like the power.t.test we saw for means.

Again, we specify 4 of the following 5 elements of the comparison, and R calculates the fifth.

- The sample size (interpreted as the # in each group, so half the total sample size)
- The true probability in group 1
- The true probability in group 2
- The significance level (α)
- The power ($1 - \beta$)

The big weakness with the power.prop.test tool is that it doesn't allow you to work with unbalanced designs.

23.6 A Revised Scenario

Suppose we can get exactly 800 subjects in total (400 sharing and 400 non-sharing). How much power would we have to detect a difference in the proportion of positive skin test results between the two groups that was identical to the data presented above or larger, using a one-sided test, with $\alpha = .10$?

```
power.prop.test(n=400, p1=.174, p2=.247, sig.level = 0.10,
                 alternative="one.sided")
```

Two-sample comparison of proportions power calculation

```
n = 400
p1 = 0.174
p2 = 0.247
sig.level = 0.1
power = 0.8954262
alternative = one.sided
```

NOTE: n is number in *each* group

We would have just under 90% power to detect such an effect.

23.7 Using the `pwr` library for Unbalanced Designs

The `pwr.2p2n.test` function in the `pwr` library can help assess the power of a test to determine a particular effect size using an unbalanced design, where n_1 is not equal to n_2 .

As before, we specify four of the following five elements of the comparison, and R calculates the fifth.

- `n1` = The sample size in group 1
- `n2` = The sample size in group 2
- `sig.level` = The significance level (α)
- `power` = The power ($1 - \beta$)
- `h` = the effect size h , which can be calculated separately in R based on the two proportions being compared: `p1` and `p2`.

23.7.1 Calculating the Effect Size `h`

To calculate the effect size for a given set of proportions, just use `ES.h(p1, p2)` which is available in the `pwr` library.

For instance, in our comparison, we have the following effect size.

```
pwr::ES.h(p1 = .174, p2 = .247)
```

```
[1] -0.1796783
```

23.8 Using `pwr.2p2n.test` in R

Suppose we can have 700 samples in group 1 (the not sharing group) but only half that many in group 2 (the group of users who share needles). How much power would we have to detect this same difference ($p_1 = .174$, $p_2 = .247$) with a 5% significance level in a two-sided test?

```
pwr::pwr.2p2n.test(h = pwr::ES.h(p1 = .174, p2 = .247),
                     n1 = 700, n2 = 350,
                     sig.level = 0.05)
```

```
difference of proportion power calculation for binomial distribution (arcsine transformation)

h = 0.1796783
n1 = 700
n2 = 350
sig.level = 0.05
power = 0.7836768
alternative = two.sided
```

NOTE: different sample sizes

Note that the headline for this output actually reads:

```
difference of proportion power calculation for binomial distribution
(arcsine transformation)
```

It appears we will have about 78% power under these circumstances.

23.8.1 Comparison to Balanced Design

How does this compare to the results with a balanced design using only 1000 drug users in total, so that we have 500 patients in each group?

```
pwr::pwr.2p2n.test(h = pwr::ES.h(p1 = .174, p2 = .247),
                     n1 = 500, n2 = 500, sig.level = 0.05)
```

```
difference of proportion power calculation for binomial distribution (arcsine transformation)

h = 0.1796783
n1 = 500
n2 = 500
sig.level = 0.05
power = 0.8108416
alternative = two.sided
```

NOTE: different sample sizes

or we could instead have used...

```
power.prop.test(p1 = .174, p2 = .247,  
                sig.level = 0.05, n = 500)
```

Two-sample comparison of proportions power calculation

```
n = 500  
p1 = 0.174  
p2 = 0.247  
sig.level = 0.05  
power = 0.8091808  
alternative = two.sided
```

NOTE: n is number in *each* group

Note that these two sample size estimation approaches are approximations, and use slightly different approaches, so it's not surprising that the answers are similar, but not completely identical.

24 Larger Contingency Tables

What will we do with tables describing data from more than two categories at a time, returning to the notion of independent (rather than paired or matched) samples? The chi-square tests we have already seen in our `twobytwo` table output will extend nicely to this scenario, especially the Pearson χ^2 (asymptotic) test.

24.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

source("data/Love-boost.R")
library(Epi)
library(janitor)
library(tidyverse)

theme_set(theme_bw())
```

We will also use some functions from the `vcd` package.

24.2 A 2x3 Table: Comparing Response to Active vs. Placebo

The table below, containing 2 rows and 3 columns of data (ignoring the marginal totals) specifies the number of patients who show *complete*, *partial*, or *no response* after treatment with either **active** medication or a **placebo**.

Group	None	Partial	Complete
Active	16	26	29
Placebo	24	26	18

Is there a statistically significant association here? That is to say, is there a statistically significant difference between the treatment groups in the distribution of responses?

24.2.1 Getting the Table into R

To answer this, we'll have to get the data from this contingency table into a matrix in R. Here's one approach...

```
T1 <- matrix(c(16,26,29,24,26,18), ncol=3, nrow=2, byrow=TRUE)
rownames(T1) <- c("Active", "Placebo")
colnames(T1) <- c("None", "Partial", "Complete")
```

```
T1
```

	None	Partial	Complete
Active	16	26	29
Placebo	24	26	18

24.2.2 Manipulating the Table's presentation

We can add margins to the matrix to get a table including row and column totals.

```
addmargins(T1)
```

	None	Partial	Complete	Sum
Active	16	26	29	71
Placebo	24	26	18	68
Sum	40	52	47	139

Instead of the counts, we can tabulate the proportion of all patients within each cell.

```
prop.table(T1)
```

	None	Partial	Complete
Active	0.1151079	0.1870504	0.2086331
Placebo	0.1726619	0.1870504	0.1294964

Now, to actually obtain a p value and perform the significance test with H_0 : rows and columns are independent vs. H_A : rows and columns are associated, we simply run a Pearson chi-square test on T1 ...

```
chisq.test(T1)
```

```
Pearson's Chi-squared test
```

```
data: T1  
X-squared = 4.1116, df = 2, p-value = 0.128
```

Thanks to a p-value of about 0.13 (using the Pearson chi-square test) our conclusion would be to retain the null hypothesis of independence in this setting.

We could have run a Fisher's exact test, too, if we needed it.

```
fisher.test(T1)
```

```
Fisher's Exact Test for Count Data
```

```
data: T1  
p-value = 0.1346  
alternative hypothesis: two.sided
```

The Fisher exact test p value is also 0.13. Either way, there is insufficient evidence to conclude that there is a (true) difference in the distributions of responses.

24.3 Accuracy of Death Certificates (A 6x3 Table)

The table below compiles data from six studies designed to investigate the accuracy of death certificates. The original citation is Kircher T, Nelson J, Burdo H (1985) The autopsy as a measure of accuracy of the death certificate. *NEJM*, 313, 1263-1269. 5373 autopsies were compared to the causes of death listed on the certificates. Of those, 3726 were confirmed to be accurate, 783 either lacked information or contained inaccuracies but did not require recoding of the underlying cause of death, and 864 were incorrect and required recoding. Do the results across studies appear consistent?

Date of Study	[Confirmed] Accurate	[Inaccurate] No Change	[Incorrect] Recoding	Total
1955-1965	2040	367	327	2734
1970	149	60	48	257
1970-1971	288	25	70	383
1975-1977	703	197	252	1152
1977-1978	425	62	88	575

Date of Study	[Confirmed] Accurate	[Inaccurate] Change	No [Incorrect] Recoding	Total
1980	121	72	79	272
Total	3726	783	864	5373

24.4 The Pearson Chi-Square Test of Independence

We can assess the homogeneity of the confirmation results (columns) we observe in the table using a Pearson chi-squared test of independence.

- The null hypothesis is that the rows and columns are independent.
- The alternative hypothesis is that there is an association between the rows and the columns.

```
death.tab <- matrix(c(2040,367,327,149,60,48,288,25,70,703,
                     197,252,425,62,88,121,72,79), byrow=TRUE, nrow=6)
rownames(death.tab) <- c("1955-65", "1970", "1970-71", "1975-77", "1977-78",
                         "1980")
colnames(death.tab) <- c("Confirmed", "Inaccurate", "Incorrect")

addmargins(death.tab)
```

	Confirmed	Inaccurate	Incorrect	Sum
1955-65	2040	367	327	2734
1970	149	60	48	257
1970-71	288	25	70	383
1975-77	703	197	252	1152
1977-78	425	62	88	575
1980	121	72	79	272
Sum	3726	783	864	5373

To see the potential heterogeneity across rows in these data, we should perhaps also look at the proportions of autopsies in each of the three accuracy categories for each study.

```
addmargins(round(100*prop.table(death.tab,1),1),2)
```

	Confirmed	Inaccurate	Incorrect	Sum
1955-65	74.6	13.4	12.0	100
1970	58.0	23.3	18.7	100

1970-71	75.2	6.5	18.3	100
1975-77	61.0	17.1	21.9	100
1977-78	73.9	10.8	15.3	100
1980	44.5	26.5	29.0	100

In three of the studies, approximately 3/4 of the results were confirmed. In the other three, 45%, 58% and 61% were confirmed. It looks like there's a fair amount of variation in results across studies. To see if this is true, formally, we run Pearson's chi-square test of independence, where the null hypothesis is that the rows and columns are independent, and the alternative hypothesis is that there is an association between the rows and the columns.

```
chisq.test(death.tab)
```

```
Pearson's Chi-squared test

data: death.tab
X-squared = 209.09, df = 10, p-value < 2.2e-16
```

The chi-square test statistic is 200 on 10 degrees of freedom, yielding $p < 0.0001$.

Autopsies are not performed at random; in fact, many are done because the cause of death listed on the certificate is uncertain. What problems may arise if you attempt to use the results of these studies to make inference about the population as a whole?

24.5 Three-Way Tables: A 2x2xK Table and a Mantel-Haenszel Analysis

The material I discuss in this section is attributable to Jeff Simonoff and his book *Analyzing Categorical Data*. The example is taken from Section 8.1 of that book.

A three-dimensional or three-way table of counts often reflects a situation where the rows and columns refer to variables whose association is of primary interest to us, and the third factor (a layer, or strata) describes a control variable, whose effect on our primary association is something we are *controlling* for in the analysis.

24.5.1 Smoking and Mortality in the UK

In the early 1970s and then again 20 years later, in Whickham, United Kingdom, surveys yielded the following relationship between whether a person was a smoker at the time of the original survey and whether they were still alive 20 years later¹.

```
whickham1 <- matrix(c(502, 230, 443, 139), byrow=TRUE, nrow=2)
rownames(whickham1) <- c("Non-Smoker", "Smoker")
colnames(whickham1) <- c("Alive", "Dead")
addmargins(whickham1)
```

	Alive	Dead	Sum
Non-Smoker	502	230	732
Smoker	443	139	582
Sum	945	369	1314

Here's the two-by-two table analysis.

```
twoby2(whickham1)
```

2 by 2 table analysis:

Outcome : Alive

Comparing : Non-Smoker vs. Smoker

	Alive	Dead	P(Alive)	95% conf. interval
Non-Smoker	502	230	0.6858	0.6512 0.7184
Smoker	443	139	0.7612	0.7248 0.7941

95% conf. interval

Relative Risk: 0.9010 0.8427 0.9633

Sample Odds Ratio: 0.6848 0.5353 0.8761

Conditional MLE Odds Ratio: 0.6850 0.5307 0.8822

Probability difference: -0.0754 -0.1230 -0.0266

Exact P-value: 0.0030

Asymptotic P-value: 0.0026

¹See Appleton et al. 1996. Ignoring a Covariate: An Example of Simpson's Paradox. The American Statistician, 50, 340-341.

There is a detectable association between smoking and mortality, but it isn't the one you might expect.

- The odds ratio is 0.68, implying that the odds of having lived were only 68% as large for non-smokers as for smokers.
- Does that mean that smoking is *good* for you?

Not likely. There is a key “lurking” variable here - a variable that is related to both smoking and mortality that is obscuring the actual relationship - namely, age.

24.5.2 The whickham data with age, too

The table below gives the mortality experience separated into subtables by initial age group.

```
age <- c(rep("18-24", 4), rep("25-34", 4),
         rep("35-44", 4), rep("45-54", 4),
         rep("55-64", 4), rep("65-74", 4),
         rep("75+", 4))

smoking <- c(rep(c("Smoker", "Smoker", "Non-Smoker", "Non-Smoker"), 7))
status <- c(rep(c("Alive", "Dead"), 14))
counts <- c(53, 2, 61, 1, 121, 3, 152, 5,
           95, 14, 114, 7, 103, 27, 66, 12,
           64, 51, 81, 40, 7, 29, 28, 101,
           0, 13, 0, 64)

whickham2 <- tibble(smoking, status, age, counts) |>
  mutate(smoking = factor(smoking),
         status = factor(status),
         age = factor(age))

whickham2

# A tibble: 28 x 4
  smoking   status   age   counts
  <fct>     <fct>   <fct>   <dbl>
1 Smoker    Alive    18-24     53
2 Smoker    Dead     18-24      2
3 Non-Smoker Alive    18-24     61
4 Non-Smoker Dead     18-24      1
5 Smoker    Alive    25-34    121
6 Smoker    Dead     25-34      3
```

```

7 Non-Smoker Alive 25-34    152
8 Non-Smoker Dead 25-34      5
9 Smoker      Alive 35-44    95
10 Smoker     Dead 35-44    14
# ... with 18 more rows
# i Use `print(n = ...)` to see more rows

whick_t2 <-
  xtabs(counts ~ smoking + status + age, data = whickham2)

whick_t2

, , age = 18-24

  status
smoking      Alive Dead
Non-Smoker    61    1
Smoker        53    2

, , age = 25-34

  status
smoking      Alive Dead
Non-Smoker    152   5
Smoker        121   3

, , age = 35-44

  status
smoking      Alive Dead
Non-Smoker    114   7
Smoker        95   14

, , age = 45-54

  status
smoking      Alive Dead
Non-Smoker    66   12
Smoker        103  27

, , age = 55-64

```

status		
smoking	Alive	Dead
Non-Smoker	81	40
Smoker	64	51

, , age = 65-74

status		
smoking	Alive	Dead
Non-Smoker	28	101
Smoker	7	29

, , age = 75+

status		
smoking	Alive	Dead
Non-Smoker	0	64
Smoker	0	13

The sample odds ratios for remaining Alive comparing Non-Smokers to Smokers for each of these subtables (except the last one, where the odds ratio is undefined because of the zero cells) are calculated using the usual cross-product ratio approach, which yields the following results.

Age Group	Odds Ratio
18-24	2.30
25-34	0.75
35-44	2.40
45-54	1.44
55-64	1.61
65-74	1.15
75+	Undefined

Thus, for all age groups except 25-34 year olds, not smoking appears to be associated with higher odds of remaining alive.

Why? Not surprisingly, there is a strong association between age and mortality, with mortality rates being very low for young people (2.5% for 18-24 year olds) and increasing to 100% for 75+ year olds.

There is also an association between age and smoking, with smoking rates peaking in the 45-54 year old range and then falling off rapidly. In particular, respondents who were 65 and older

at the time of the first survey had very low smoking rates (25.4%) but very high mortality rates (85.5%). Smoking was hardly the cause, however, since even among the 65-74 year olds mortality was higher among smokers (80.6%) than it was among non-smokers (78.3%). A flat version of the table (`ftable` in R) can help us with these calculations.

```
ftable(whick_t2)
```

		age	18-24	25-34	35-44	45-54	55-64	65-74	75+
smoking	status								
Non-Smoker	Alive		61	152	114	66	81	28	0
	Dead		1	5	7	12	40	101	64
Smoker	Alive		53	121	95	103	64	7	0
	Dead		2	3	14	27	51	29	13

24.5.3 Checking Assumptions: The Woolf test

We can also obtain a test (using the `woolf_test` function, in the `vcd` library) to see if the common odds ratio estimated in the Mantel-Haenszel procedure is reasonable for all age groups. In other words, the Woolf test is a test of the assumption of homogeneous odds ratios across the six age groups.

If the Woolf test is significant, it suggests that the Cochran-Mantel-Haenszel test is not appropriate, since the odds ratios for smoking and mortality vary too much in the sub-tables by age group. Here, we have the following log odds ratios (estimated using conditional maximum likelihood, rather than cross-product ratios) and the associated Woolf test.

```
## Next two results use the vcd package
vcdd::oddsratio(whick_t2, log = TRUE)
```

log odds ratios for smoking and status by age

18-24	25-34	35-44	45-54	55-64	65-74
0.65018114	-0.22473479	0.84069420	0.34608770	0.47421763	0.09933253
75+					
-1.56397554					

```
vcdd::woolf_test(whick_t2)
```

Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)

```
data: whick_t2
X-squared = 3.2061, df = 6, p-value = 0.7826
```

As you can see, the Woolf test is not close to our usual standards for statistically detectable results, implying the common odds ratio is at least potentially reasonable for all age groups (or at least the ones under ages 75, where some data are available.)

24.5.4 The Cochran-Mantel-Haenszel Test

So, the marginal table looking at smoking and mortality combining all age groups isn't the most meaningful summary of the relationship between smoking and mortality. Instead, we need to look at the *conditional* association of smoking and mortality, **given age**, to address our interests.

The null hypothesis would be that, in the population, smoking and mortality are independent within strata formed by age group. In other words, H_0 requires that smoking be of no value in predicting mortality once age has been accounted for.

The alternative hypothesis would be that, in the population, smoking and mortality are associated within the strata formed by age group. In other words, H_A requires that smoking be of at least some value in predicting mortality even after age has been accounted for.

We can consider the evidence that helps us choose between these two hypotheses with a Cochran-Mantel-Haenszel test, which is obtained in R through the `mantelhaen.test` function. This test requires us to assume that, in the population and within each age group, the smoking-mortality odds ratio is the same. Essentially, this means that the association of smoking with mortality is the same for older and younger people.

```
mantelhaen.test(whick_t2, conf.level = 0.90)
```

```
Mantel-Haenszel chi-squared test with continuity correction

data: whick_t2
Mantel-Haenszel X-squared = 5.435, df = 1, p-value = 0.01974
alternative hypothesis: true common odds ratio is not equal to 1
90 percent confidence interval:
 1.143198 2.041872
sample estimates:
common odds ratio
 1.52783
```

- The Cochran-Mantel-Haenszel test statistic is a bit larger than 5 (after a continuity correction) leading to a p value of 0.02, indicating strong rejection of the null hypothesis of conditional independence of smoking and survival given age.
- The estimated common conditional odds ratio is 1.53. This implies that (adjusting for age) being a non-smoker is associated with 53% higher odds of being alive 20 years later than being a smoker.
- A 90% confidence interval for that common odds ratio is (1.14, 2.04), reinforcing rejection of the null hypothesis of conditional independence (where the odds ratio would be 1).

24.5.5 Without the Continuity Correction

By default, R presents the Mantel-Haenszel test with a continuity correction, when used for a 2x2xK table. In virtually all cases, go ahead and do this, but as you can see below, the difference it makes in this case is modest.

```
mantelhaen.test(whick_t2, correct=FALSE, conf.level = 0.90)
```

```
Mantel-Haenszel chi-squared test without continuity correction

data: whick_t2
Mantel-Haenszel X-squared = 5.8443, df = 1, p-value = 0.01563
alternative hypothesis: true common odds ratio is not equal to 1
90 percent confidence interval:
 1.143198 2.041872
sample estimates:
common odds ratio
 1.52783
```

25 Analysis of Variance

25.1 Setup: Packages Used Here

```
knitr::opts_chunk$set(comment = NA)

library(tidyverse)

theme_set(theme_bw())
```

We will also use functions from the `ggridges` and `mosaic` packages.

25.2 National Youth Fitness Survey

Recall the National Youth Fitness Survey, which we explored a small piece of in some detail earlier in these notes. We'll look at a different part of the same survey here - specifically the 280 children whose data are captured in the `nyfs2` file.

```
nyfs2 <- read_csv("data/nyfs2.csv", show_col_types = FALSE)

nyfs2

# A tibble: 280 x 21
  subje~1 sex    age.e~2 race.~3 english incom~4 incom~5 inc.t~6 weigh~7 heigh~8
  <dbl> <chr>   <dbl> <chr>     <dbl> <chr>     <dbl> <chr>     <dbl> <dbl> <dbl>
1 73228 Male      4 5 Othe~      1 Low (b~ 0 to 4~    0       17    104.
2 72393 Male      4 2 Non--     1 Low (b~ 0 to 4~    0       16.4   102.
3 73303 Male      3 2 Non--     1 Low (b~ 0 to 4~   0.16    16.4   98.7
4 72786 Male      5 1 Non--     1 Low (b~ 0 to 4~   0.04    19.1   114.
5 73048 Male      3 2 Non--     1 Low (b~ 0 to 4~   0.17    14     93.8
6 72556 Fema~     4 2 Non--     1 Low (b~ 0 to 4~   0.06    20.8   106.
7 72580 Fema~     5 2 Non--     1 Low (b~ 0 to 4~   0.17    18     107.
8 72532 Fema~     4 4 Othe~     0 Low (b~ 0 to 4~   0.02    16.7   101.
```

```

9   73012 Male      4 1 Non--      1 Low (b~ 0 to 4~    0.21    19.8    106.
10  72099 Male      6 1 Non--      1 Low (b~ 0 to 4~    0.2     31.4    121.
# ... with 270 more rows, 11 more variables: bmi <dbl>, bmi.group <dbl>,
#   bmi.cat <chr>, arm.length <dbl>, arm.circ <dbl>, waist.circ <dbl>,
#   calf.circ <dbl>, calf.skinfold <dbl>, triceps.skinfold <dbl>,
#   subscap.skinfold <dbl>, GMQ <dbl>, and abbreviated variable names
#   1: subject.id, 2: age.exam, 3: race.eth, 4: income.cat3, 5: income.detail,
#   6: inc.to.pov, 7: weight.kg, 8: height.cm
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names

```

25.3 Comparing Gross Motor Quotient Scores by Income Level (3 Categories)

```

nyfs2a <- nyfs2 |>
  select(subject.id, income.cat3, GMQ) |>
  arrange(subject.id)

```

In this first analysis, we'll compare the population mean on the Gross Motor Quotient evaluation of these kids across three groups defined by income level. Higher values of this GMQ measure indicate improved levels of gross motor development, both in terms of locomotor and object control. See https://www.cdc.gov/Nchs/Nyfs/Y_GMX.htm for more details.

```

nyfs2a |>
  group_by(income.cat3) |>
  summarise(n = n(), mean(GMQ), median(GMQ))

# A tibble: 3 x 4
  income.cat3       n `mean(GMQ)` `median(GMQ)`
  <chr>        <int>      <dbl>       <dbl>
1 High (65K or more)    92      95.7        97
2 Low (below 25K)      98      97.0        97
3 Middle (25 - 64K)    90      95.4        94

```

Uh, oh. We should rearrange those income categories to match a natural order from low to high.

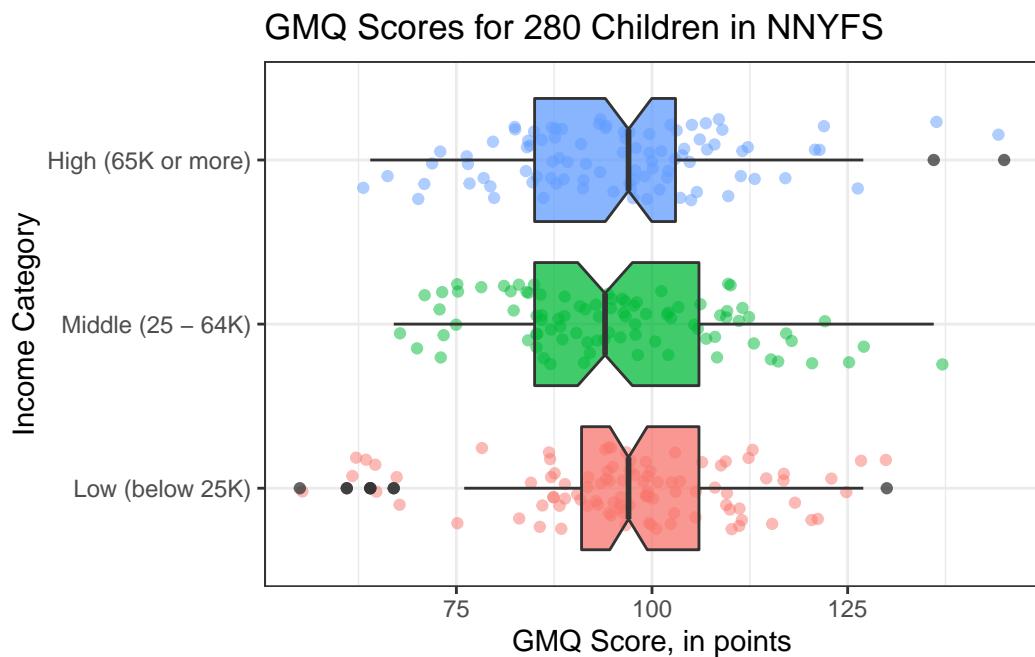
```

nyfs2a$income.cat3 <-
  fct_relevel(nyfs2a$income.cat3,
              "Low (below 25K)", "Middle (25 - 64K)", "High (65K or more)")

```

When working with three independent samples, I use graphs analogous to those we built for two independent samples.

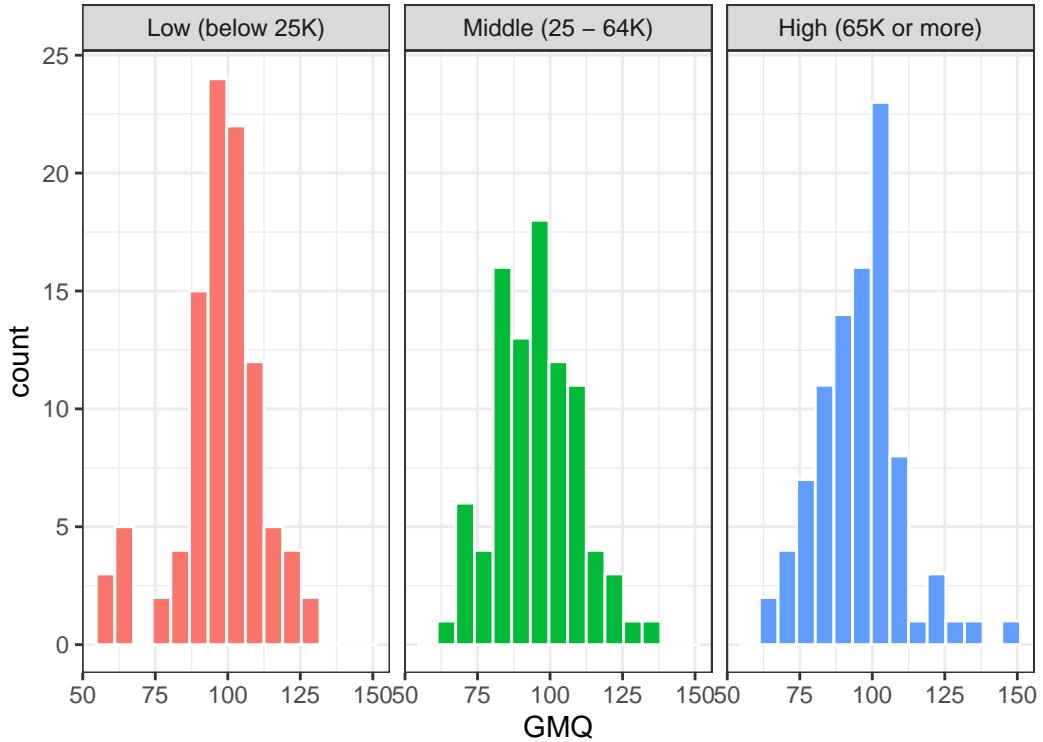
```
ggplot(nyfs2a, aes(x = income.cat3, y = GMQ, fill = income.cat3)) +
  geom_jitter(aes(color = income.cat3), alpha = 0.5, width = 0.25) +
  geom_boxplot(notch = TRUE, alpha = 0.75) +
  coord_flip() +
  guides(fill = "none", col = "none") +
  labs(title = "GMQ Scores for 280 Children in NNYFS",
       y = "GMQ Score, in points", x = "Income Category")
```



In addition to this comparison boxplot, we might consider faceted plots, like these histograms.

```
ggplot(nyfs2a, aes(x = GMQ, fill = income.cat3)) +
  geom_histogram(bins = 15, col = "white") +
  guides(fill = FALSE) +
  facet_wrap(~ income.cat3)
```

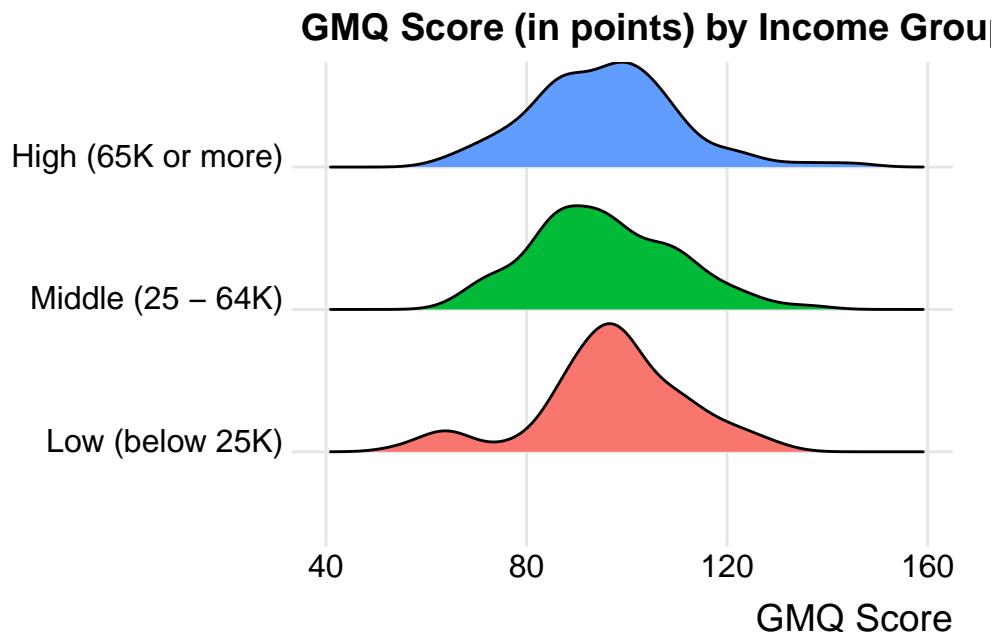
Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.



Or, if we want to ignore the (modest) sample size differences, we might consider density functions, perhaps through a ridgeline plot.

```
ggplot(nyfs2a, aes(x = GMQ, y = income.cat3, fill = income.cat3)) +
  ggridges::geom_density_ridges(scale = 0.9) +
  guides(fill = FALSE) +
  labs(title = "GMQ Score (in points) by Income Group",
       x = "GMQ Score", y = "") +
  ggridges::theme_ridges()
```

Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.



```
by(nyfs2a$GMQ, nyfs2a$income.cat3, mosaic::favstats)
```

```
nyfs2a$income.cat3: Low (below 25K)
min Q1 median Q3 max      mean      sd n missing
 55  91     97 106 130 97.03061 14.79444 98      0
```

```
nyfs2a$income.cat3: Middle (25 - 64K)
min Q1 median Q3 max      mean      sd n missing
 67  85     94 106 136 95.36667 14.15123 90      0
```

```
nyfs2a$income.cat3: High (65K or more)
min Q1 median Q3 max      mean      sd n missing
 64  85     97 103 145 95.72826 14.49525 92      0
```

25.4 Alternative Procedures for Comparing More Than Two Means

Now, if we only had two independent samples, we'd be choosing between a pooled t test, a Welch t test, and a non-parametric procedure like the Wilcoxon-Mann-Whitney rank sum test, or even perhaps a bootstrap alternative.

In the case of more than two independent samples, we have methods analogous to the Welch test, and the rank sum test, and even the bootstrap, but we're going to be far more likely to select the **analysis of variance** (ANOVA) or an equivalent regression-based approach. These are the extensions of the pooled t test. Unless the sample outcome data are very clearly not Normally distributed, and no transformation is available which makes them appear approximately Normal in all of the groups we are comparing, we will stick with ANOVA.

25.4.1 Extending the Welch Test to > 2 Independent Samples

It is possible to extend the Welch two-sample t test (not assuming equal population variances) into an analogous one-factor analysis for comparing population means based on independent samples from more than two groups.

If we want to compare the population mean GMQ levels across those three income groups without assuming equal population variances, `oneway.test` is up to the task. The hypotheses being tested here are:

- H₀: All three means are the same vs.
- H_A: At least one of the population means is different than the others.

```
oneway.test(GMQ ~ income.cat3, data = nyfs2a)
```

```
One-way analysis of means (not assuming equal variances)

data: GMQ and income.cat3
F = 0.3416, num df = 2.00, denom df = 184.41, p-value = 0.7111
```

We get a p value, but this isn't much help, though, because we don't have any measure of effect size, nor do we have any confidence intervals. Like the analogous Welch t test, this approach allows us to forego the assumption of equal population variances in each of the three income groups, but it still requires us to assume that the populations are Normally distributed.

That said, most of the time when we have more than two levels of the factor of interest, we won't bother worrying about the equal population variance assumption, and will just use the one-factor ANOVA approach (with pooled variances) described below, to make the comparisons of interest.

25.4.2 Extending the Rank Sum Test to > 2 Independent Samples

It is also possible to extend the Wilcoxon-Mann-Whitney two-sample test into an analogous one-factor analysis called the **Kruskal-Wallis test** for comparing population measures of location based on independent samples from more than two groups.

If we want to compare the centers of the distributions of population GMQ score across our three income groups without assuming Normality, we can use `kruskal.test`.

The hypotheses being tested here are still as before, but for a measure of location other than the population mean

```
kruskal.test(GMQ ~ income.cat3, data = nyfs2a)
```

```
Kruskal-Wallis rank sum test

data: GMQ by income.cat3
Kruskal-Wallis chi-squared = 2.3202, df = 2, p-value = 0.3135
```

Again, note that this isn't much help, though, because we don't have any measure of effect size, nor do we have any confidence intervals.

That said, most of the time when we have more than two levels of the factor of interest, we won't bother worrying about potential violations of the Normality assumption unless they are glaring, and will just use the usual one-factor ANOVA approach (with pooled variances) described below, to make the comparisons of interest.

25.4.3 Can we use the bootstrap to compare more than two means?

Sure. There are both ANOVA and ANCOVA analogues using the bootstrap, and in fact, there are power calculations based on the bootstrap, too. If you want to see some example code, look at <https://sammancuso.com/2017/11/01/model-based-bootstrapped-anova-and-ancova/>

25.5 The Analysis of Variance

Extending the two-sample t test (assuming equal population variances) into a comparison of more than two samples uses the **analysis of variance** or ANOVA.

This is an analysis of a continuous outcome variable on the basis of a single categorical factor, in fact, it's often called one-factor ANOVA or one-way ANOVA to indicate that the outcome is being split up into the groups defined by a single factor.

The null hypothesis is that the population means are all the same, and the alternative is that this is not the case. When there are just two groups, then this boils down to an F test that is equivalent to the Pooled t test.

25.5.1 The `oneway.test` approach

R will produce some elements of a one-factor ANOVA using the `oneway.test` command:

```
oneway.test(GMQ ~ income.cat3, data = nyfs2a, var.equal=TRUE)
```

```
One-way analysis of means

data: GMQ and income.cat3
F = 0.34687, num df = 2, denom df = 277, p-value = 0.7072
```

This isn't the full analysis, though, which would require a more complete ANOVA table. There are two equivalent approaches to obtaining the full ANOVA table when comparing a series of 2 or more population means based on independent samples.

25.5.2 Using the `aov` approach and the `summary` function

Here's one possible ANOVA table, which doesn't require directly fitting a linear model.

```
summary(aov(GMQ ~ income.cat3, data = nyfs2a))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income.cat3	2	146	72.85	0.347	0.707
Residuals	277	58174	210.01		

25.5.3 Using the `anova` function after fitting a linear model

An equivalent way to get identical results in a slightly different format runs the linear model behind the ANOVA approach directly.

```
anova(lm(GMQ ~ income.cat3, data = nyfs2a))
```

Analysis of Variance Table

Response: GMQ

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income.cat3	2	146	72.848	0.3469	0.7072
Residuals	277	58174	210.014		

25.6 Interpreting the ANOVA Table

25.6.1 What are we Testing?

The null hypothesis for the ANOVA table is that the population means of the outcome across the various levels of the factor of interest are all the same, against a two-sided alternative hypothesis that the level-specific population means are not all the same.

Specifically, if we have a grouping factor with k levels, then we are testing:

- H₀: All k population means are the same.
- H_A: At least one of the population means is different from the others.

25.6.2 Elements of the ANOVA Table

The ANOVA table breaks down the variation in the outcome explained by the k levels of the factor of interest, and the variation in the outcome which remains (the Residual, or Error).

Specifically, the elements of the ANOVA table are:

1. the degrees of freedom (labeled Df) for the factor of interest and for the Residuals
2. the sums of squares (labeled Sum Sq) for the factor of interest and for the Residuals
3. the mean square (labeled Mean Sq) for the factor of interest and for the Residuals
4. the ANOVA F test statistic (labeled F value), which is used to generate
5. the p value for the comparison assessed by the ANOVA model, labeled Pr(>F)

25.6.3 The Degrees of Freedom

```
anova(lm(GMQ ~ income.cat3, data = nyfs2a))
```

Analysis of Variance Table

Response: GMQ

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income.cat3	2	146	72.848	0.3469	0.7072
Residuals	277	58174	210.014		

- The **degrees of freedom** attributable to the factor of interest (here, Income category) is the number of levels of the factor minus 1. Here, we have three Income categories (levels), so $df(income.cat3) = 2$.
- The total degrees of freedom are the number of observations (across all levels of the factor) minus 1. We have 280 GMQ scores in the `nyfs2a` data, so the $df(Total)$ must be 279, although the Total row isn't shown by R in its output.
- The Residual degrees of freedom are the Total df - Factor df. So, here, that's $279 - 2 = 277$.

25.6.4 The Sums of Squares

```
anova(lm(GMQ ~ income.cat3, data = nyfs2a))
```

Analysis of Variance Table

Response: GMQ

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income.cat3	2	146	72.848	0.3469	0.7072
Residuals	277	58174	210.014		

- The sum of squares (often abbreviated SS or Sum Sq) represents variation explained.
- The factor SS is the sum across all levels of the factor of the sample size for the level multiplied by the squared difference between the level mean and the overall mean across all levels. Here, $SS(income.cat3) = 146$
- The total SS is the sum across all observations of the square of the difference between the individual values and the overall mean. Here, that is $146 + 58174 = 58320$
- Residual SS = Total SS - Factor SS.
- Also of interest is a calculation called η^2 , ("eta-squared"), which is equivalent to R^2 in a linear model.
 - $SS(\text{Factor}) / SS(\text{Total}) =$ the proportion of variation in our outcome (here, GMQ) explained by the variation between groups (here, income groups)
 - In our case, $\eta^2 = 146 / (146 + 58174) = 146 / 58320 = 0.0025$
 - So, Income Category alone accounts for about 0.25% of the variation in GMQ levels observed in these data.

25.6.5 The Mean Square

```
anova(lm(GMQ ~ income.cat3, data = nyfs2a))
```

Analysis of Variance Table

Response: GMQ

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income.cat3	2	146	72.848	0.3469	0.7072
Residuals	277	58174	210.014		

- The Mean Square is the Sum of Squares divided by the degrees of freedom, so $MS(\text{Factor}) = SS(\text{Factor})/\text{df}(\text{Factor})$.
- In our case, $MS(\text{income.cat3}) = SS(\text{income.cat3})/\text{df}(\text{income.cat3}) = 146 / 2 = 72.848$ (notice that R maintains more decimal places than it shows for these calculations) and
- $MS(\text{Residuals}) = SS(\text{Residuals}) / \text{df}(\text{Residuals}) = 58174 / 277 = 210.014$.
 - $MS(\text{Residuals})$ or $MS(\text{Error})$ is an estimate of the residual variance which corresponds to σ^2 in the underlying linear model for the outcome of interest, here GMQ.

25.6.6 The F Test Statistic and p Value

```
anova(lm(GMQ ~ income.cat3, data = nyfs2a))
```

Analysis of Variance Table

Response: GMQ

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income.cat3	2	146	72.848	0.3469	0.7072
Residuals	277	58174	210.014		

- The ANOVA F test is obtained by calculating $MS(\text{Factor}) / MS(\text{Residuals})$. So in our case, $F = 72.848 / 210.014 = 0.3469$
- The F test statistic is then compared to a specific F distribution to obtain a p value, which is shown here to be 0.7072
- Specifically, the observed F test statistic is compared to an F distribution with numerator df = Factor df, and denominator df = Residual df to obtain the p value.
 - Here, we have $SS(\text{Factor}) = 146$ (approximately), and $\text{df}(\text{Factor}) = 2$, leaving $MS(\text{Factor}) = 72.848$

- We have $SS(\text{Residual}) = 58174$, and $df(\text{Residual}) = 277$, leaving $MS(\text{Residual}) = 210.014$
- $MS(\text{Factor}) / MS(\text{Residual}) = F \text{ value} = 0.3469$, which, when compared to an F distribution with 2 and 277 degrees of freedom, yields a p value of 0.7072

25.7 The Residual Standard Error

The residual standard error is simply the square root of the variance estimate $MS(\text{Residual})$. Here, $MS(\text{Residual}) = 210.014$, so the Residual standard error = 14.49 points.

25.8 The Proportion of Variance Explained by the Factor

We will often summarize the proportion of the variation explained by the factor. The summary statistic is called eta-squared (η^2), and is equivalent to the R^2 value we have seen previously in linear regression models.

Again, $\eta^2 = SS(\text{Factor}) / SS(\text{Total})$

Here, we have - $SS(\text{income.cat3}) = 146$ and $SS(\text{Residuals}) = 58174$, so $SS(\text{Total}) = 58320$ - Thus, $\eta^2 = SS(\text{Factor})/SS(\text{Total}) = 146/58320 = 0.0025$

The income category accounts for 0.25% of the variation in GMQ levels: only a tiny fraction.

25.9 The Regression Approach to Compare Population Means based on Independent Samples

This approach is equivalent to the ANOVA approach, and thus also (when there are just two samples to compare) to the pooled-variance t test. We run a linear regression model to predict the outcome (here, GMQ) on the basis of the categorical factor with three levels (here, `income.cat3`)

```
summary(lm(GMQ ~ income.cat3, data=nyfs2a))
```

```
Call:
lm(formula = GMQ ~ income.cat3, data = nyfs2a)
```

```
Residuals:
    Min     1Q Median     3Q    Max

```

```

-42.031 -9.031 -0.031 8.969 49.272

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 97.031     1.464   66.282 <2e-16 ***
income.cat3Middle (25 - 64K) -1.664     2.116  -0.786  0.432
income.cat3High (65K or more) -1.302     2.104  -0.619  0.536
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.49 on 277 degrees of freedom
Multiple R-squared: 0.002498, Adjusted R-squared: -0.004704
F-statistic: 0.3469 on 2 and 277 DF, p-value: 0.7072

```

25.9.1 Interpreting the Regression Output

This output tells us many things, but for now, we'll focus just on the coefficients output, which tells us that:

- the point estimate for the population mean GMQ score across “Low” income subjects is 97.03
- the point estimate (sample mean difference) for the difference in population mean GMQ level between the “Middle” and “Low” income subjects is -1.66 (in words, the Middle income kids have lower GMQ scores than the Low income kids by 1.66 points on average.)
- the point estimate (sample mean difference) for the difference in population mean GMQ level between the “High” and “Low” income subjects is -1.30 (in words, the High income kids have lower GMQ scores than the Low income kids by 1.30 points on average.)

Of course, we knew all of this already from a summary of the sample means.

```

nyfs2a |>
  group_by(income.cat3) |>
  summarise(n = n(), mean(GMQ))

# A tibble: 3 x 3
  income.cat3      n `mean(GMQ)`
  <fct>        <int>     <dbl>
1 Low (below 25K)    98      97.0
2 Middle (25 - 64K)   90      95.4
3 High (65K or more)  92      95.7

```

The model for predicting GMQ is based on two binary (1/0) indicator variables, specifically, we have:

- Estimated GMQ = 97.03 - 1.66 x [1 if Middle income or 0 if not] - 1.30 x [1 if High income or 0 if not]

The coefficients section also provides a standard error and t statistic and two-sided p value for each coefficient.

25.9.2 The Full ANOVA Table

To see the full ANOVA table corresponding to any linear regression model, we run...

```
anova(lm(GMQ ~ income.cat3, data=nyfs2a))
```

Analysis of Variance Table

```
Response: GMQ
          Df Sum Sq Mean Sq F value Pr(>F)
income.cat3    2   146   72.848  0.3469 0.7072
Residuals   277 58174 210.014
```

25.9.3 ANOVA Assumptions

The assumptions behind analysis of variance are the same as those behind a linear model. Of specific interest are:

- The samples obtained from each group are independent.
- Ideally, the samples from each group are a random sample from the population described by that group.
- In the population, the variance of the outcome in each group is equal. (This is less of an issue if our study involves a balanced design.)
- In the population, we have Normal distributions of the outcome in each group.

Happily, the F test is fairly robust to violations of the Normality assumption.

25.10 Equivalent approach to get ANOVA Results

```
summary(aov(GMQ ~ income.cat3, data = nyfs2a))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income.cat3	2	146	72.85	0.347	0.707
Residuals	277	58174	210.01		

So which of the pairs of means are driving the differences we see?

25.11 The Problem of Multiple Comparisons

1. Suppose we compare High to Low, using a test with $\alpha = 0.05$
2. Then we compare Middle to Low on the same outcome, also using $\alpha = 0.05$
3. Then we compare High to Middle, also with $\alpha = 0.05$

What is our overall α level across these three comparisons?

- It could be as bad as $0.05 + 0.05 + 0.05$, or 0.15.
- Rather than our nominal 95% confidence, we have something as low as 85% confidence across this set of simultaneous comparisons.

25.11.1 The Bonferroni solution

1. Suppose we compare High to Low, using a test with $\alpha = 0.05/3$
2. Then we compare Middle to Low on the same outcome, also using $\alpha = 0.05/3$
3. Then we compare High to Middle, also with $\alpha = 0.05/3$

Then across these three comparisons, our overall α can be (at worst)

- $0.05/3 + 0.05/3 + 0.05/3 = 0.05$
- So by changing our nominal confidence level from 95% to 98.333% in each comparison, we wind up with at least 95% confidence across this set of simultaneous comparisons.
- This is a conservative (worst case) approach.

Goal: Simultaneous comparisons of White vs AA, AA vs Other and White vs Other

```
pairwise.t.test(nyfs2a$GMQ, nyfs2a$income.cat3, p.adjust="bonferroni")
```

Pairwise comparisons using t tests with pooled SD

```
data: nyfs2a$GMQ and nyfs2a$income.cat3
```

	Low (below 25K)	Middle (25 - 64K)
Middle (25 - 64K)	1	-
High (65K or more)	1	1

P value adjustment method: bonferroni

These p values are very large.

25.11.2 Pairwise Comparisons using Tukey's HSD Method

Goal: Simultaneous (less conservative) confidence intervals and p values for our three pairwise comparisons (High vs. Low, High vs. Middle, Middle vs. Low)

```
TukeyHSD(aov(GMQ ~ income.cat3, data = nyfs2a))
```

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

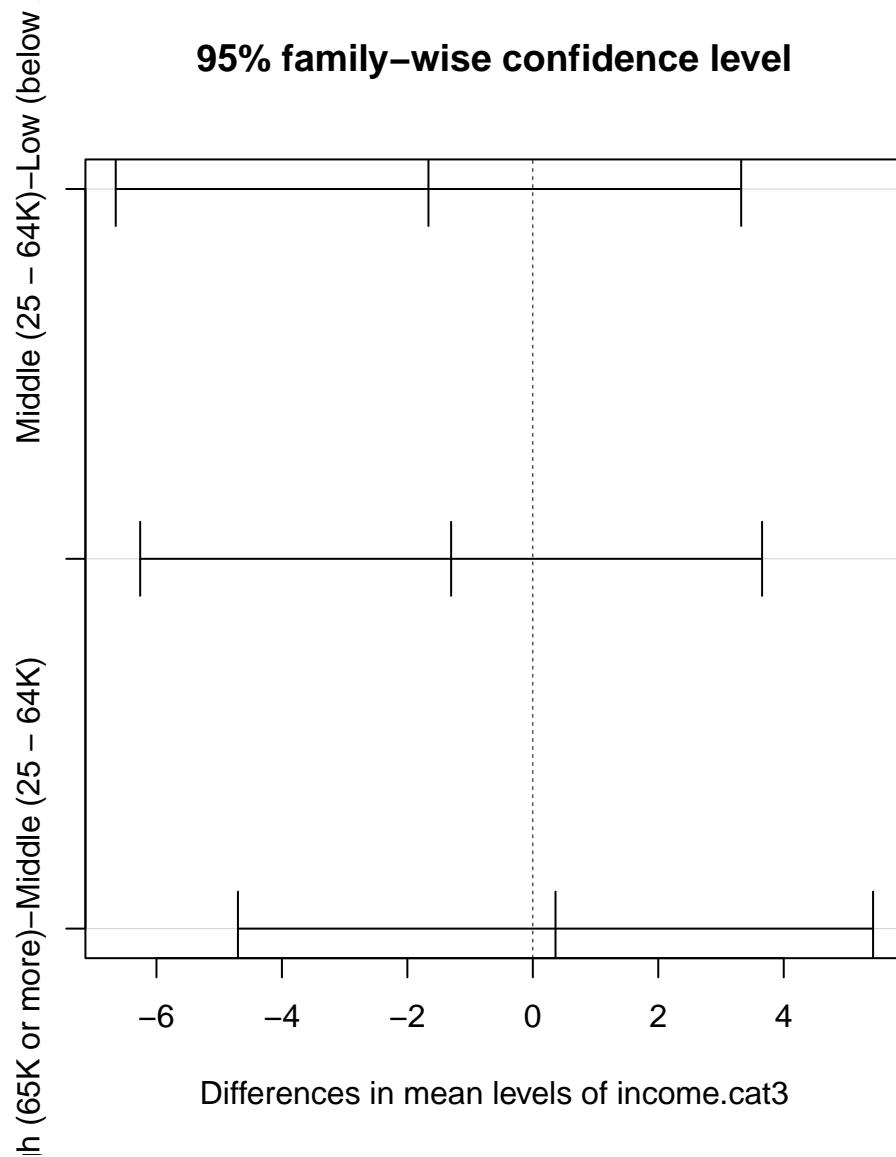
Fit: aov(formula = GMQ ~ income.cat3, data = nyfs2a)

\$income.cat3

	diff	lwr	upr	p adj
Middle (25 - 64K)-Low (below 25K)	-1.6639456	-6.649518	3.321627	0.7116745
High (65K or more)-Low (below 25K)	-1.3023514	-6.259595	3.654892	0.8098084
High (65K or more)-Middle (25 - 64K)	0.3615942	-4.701208	5.424396	0.9845073

25.11.3 Plotting the Tukey HSD results

```
plot(TukeyHSD(aov(GMQ ~ income.cat3, data = nyfs2a)))
```

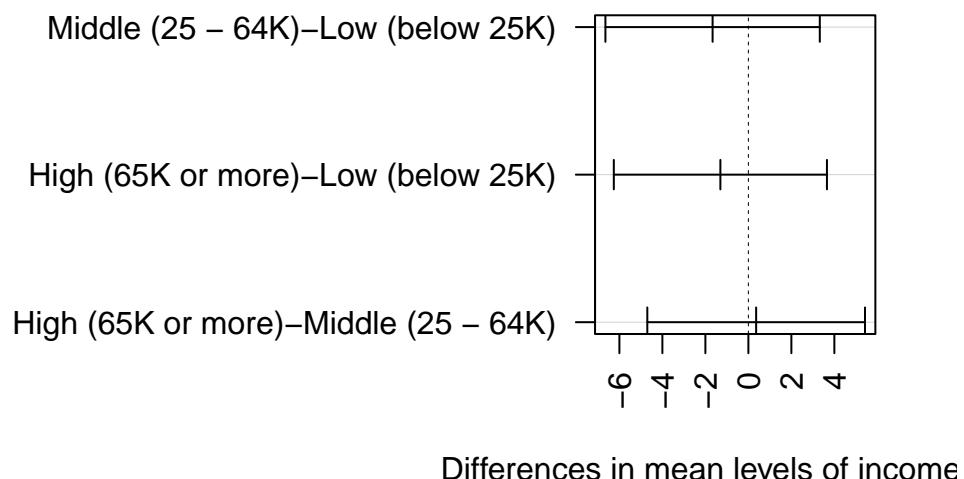


Note that the default positioning of the y axis in the plot of Tukey HSD results can be problematic. If we have longer names, in particular, for the levels of our factor, R will leave out some of the labels. We can alleviate that problem either by using the `fct_recode` function in the `forcats` package to rename the factor levels, or we can use the following code to reconfigure the margins of the plot.

```
mar.default <- c(5,6,4,2) + 0.1 # save default plotting margins
```

```
par(mar = mar.default + c(0, 12, 0, 0))
plot(TukeyHSD(aov(GMQ ~ income.cat3, data = nyfs2a)), las = 2)
```

95% family-wise confidence I



```
par(mar = mar.default) # return to normal plotting margins
```

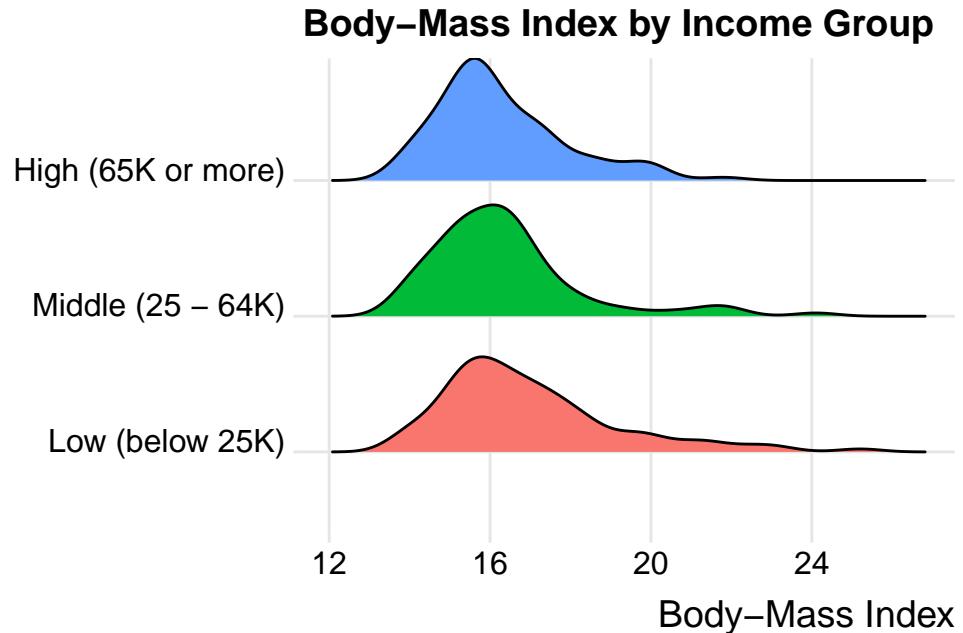
25.12 What if we consider another outcome, BMI?

We'll look at the full data set in `nyfs2` now, so we can look at BMI as a function of income.

```
nyfs2$income.cat3 <-
  fct_relevel(nyfs2$income.cat3,
              "Low (below 25K)", "Middle (25 – 64K)", "High (65K or more)")

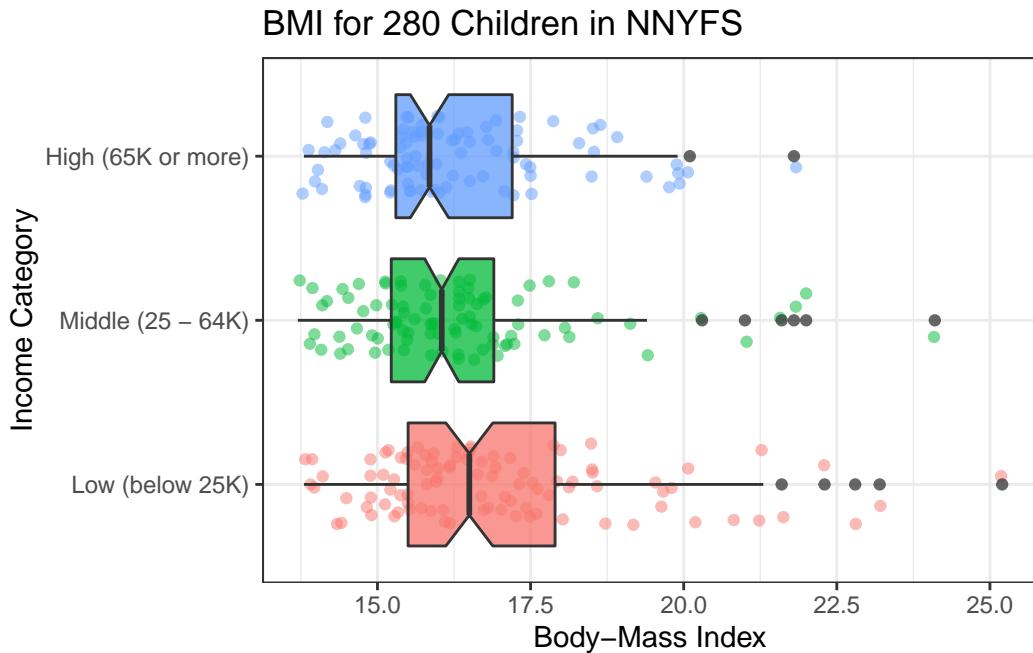
ggplot(nyfs2, aes(x = bmi, y = income.cat3, fill = income.cat3)) +
  ggridges::geom_density_ridges(scale = 0.9) +
  guides(fill = FALSE) +
  labs(title = "Body-Mass Index by Income Group",
       x = "Body-Mass Index", y = "") +
  ggridges::theme_ridges()
```

Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.



```
ggplot(nyfs2, aes(x = income.cat3, y = bmi, fill = income.cat3)) +  
  geom_jitter(aes(color = income.cat3), alpha = 0.5, width = 0.25) +  
  geom_boxplot(notch = TRUE, alpha = 0.75) +  
  theme_bw() +  
  coord_flip() +  
  guides(fill = FALSE, col = FALSE) +  
  labs(title = "BMI for 280 Children in NNYFS",  
       y = "Body-Mass Index", x = "Income Category")
```

Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> = "none")` instead.



Here are the descriptive numerical summaries:

```
mosaic::favstats(bmi ~ income.cat3, data = nyfs2)
```

	income.cat3	min	Q1	median	Q3	max	mean	sd	n	missing
1	Low (below 25K)	13.8	15.500	16.50	17.9	25.2	16.98163	2.194574	98	0
2	Middle (25 - 64K)	13.7	15.225	16.05	16.9	24.1	16.37111	1.898920	90	0
3	High (65K or more)	13.8	15.300	15.85	17.2	21.8	16.27065	1.614395	92	0

Here is the ANOVA table.

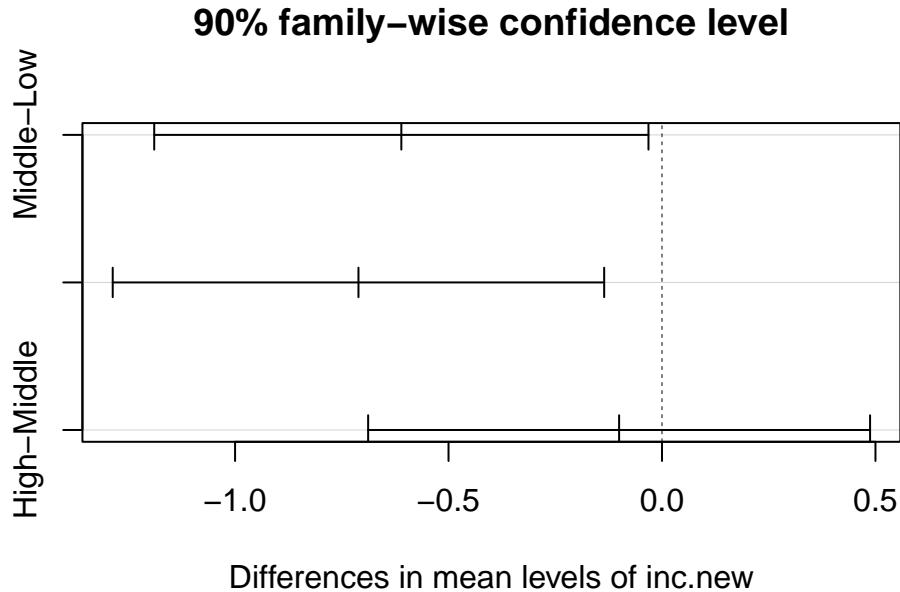
```
anova(lm(bmi ~ income.cat3, data = nyfs2))
```

Analysis of Variance Table

```
Response: bmi
          Df  Sum Sq Mean Sq F value    Pr(>F)
income.cat3   2  28.32 14.1583  3.8252 0.02298 *
Residuals  277 1025.26  3.7013
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's consider the Tukey HSD results. First, we'll create a factor with shorter labels.

```
nyfs2$inc.new <-  
  fct_recode(nyfs2$income.cat3,  
             "Low" = "Low (below 25K)", "Middle" = "Middle (25 - 64K)",  
             "High" = "High (65K or more)")  
  
plot(TukeyHSD(aov(bmi ~ inc.new, data = nyfs2),  
               conf.level = 0.90))
```



It appears that there is a detectable difference between the `bmi` means of the “Low” group and both the “High” and “Middle” group at the 90% confidence level, but no detectable difference between “Middle” and “High.” Details of those confidence intervals for those pairwise comparisons follow.

```
TukeyHSD(aov(bmi ~ inc.new, data = nyfs2),  
          conf.level = 0.90)
```

```
Tukey multiple comparisons of means  
90% family-wise confidence level
```

```
Fit: aov(formula = bmi ~ inc.new, data = nyfs2)
```

```
$inc.new
      diff      lwr      upr     p adj
Middle-Low -0.6105215 -1.1893722 -0.03167084 0.0775491
High-Low    -0.7109805 -1.2865420 -0.13541892 0.0306639
High-Middle -0.1004589 -0.6882764  0.48735849 0.9339289
```

Part III

Part C. Building Models

26 Multiple Regression: Introduction

In Chapter @ref(Hydrate-Study) in working with a study of dehydration recovery in children, we discussed many of the fundamental ideas of multiple regression. There, we provided code and insight into the scatterplot and the scatterplot matrix, fit linear models and plotted the coefficients, analyzed summary output from `summary`, `tidy` and `glance` as well as the ANOVA table, and plotted residuals vs. fitted values.

In the remaining chapters, we will build on that foundation in three additional examples.

- The `wcgs` data from the Western Collaborative Group Study, which we described and studied back in Chapter @ref(WCGS-Study).
- The `emp_bmi` data from a study published in *BMJ Open* of a nationally representative sample of over 7000 participants in the Korean Longitudinal Study of Aging.
- The `gala` data, which describe features of the 30 Galapagos Islands.

26.1 Reminders of a few Key Concepts

1. **Scatterplots** We have often accompanied our scatterplots with regression lines estimated by the method of least squares, and by loess smooths which permit local polynomial functions to display curved relationships, and occasionally presented in the form of a scatterplot matrix to enable simultaneous comparisons of multiple two-way associations.
2. **Measures of Correlation/Association** By far the most commonly used is the Pearson correlation, which is a unitless (scale-free) measure of bivariate linear association for the variables X and Y, symbolized by r , and ranging from -1 to +1. The Pearson correlation is a function of the slope of the least squares regression line, divided by the product of the standard deviations of X and Y. We have also mentioned the *Spearman* rank correlation coefficient, which is obtained by using the usual formula for a Pearson correlation, but on the ranks (1 = minimum, n = maximum, with average ranks are applied to the ties) of the X and Y values. This approach (running a correlation of the orderings of the data) substantially reduces the effect of outliers. The result still ranges from -1 to +1, with 0 indicating no linear association.
3. **Fitting Linear Models** We have fit several styles of linear model to date, including both *simple* regressions, where our outcome Y is modeled as a linear function of a single predictor X, and *multiple* regression models, where more than one predictor is used.

Functions from the `broom` package can be used to yield several crucial results, in addition to those we can obtain with a `summary` of the model. These include:

- (from `tidy`) the estimated coefficients (intercept and slope(s)) of the fitted model, and
- (from `glance`) the R^2 or coefficient of determination, which specifies the proportion of variation in our outcome accounted for by the linear model, and various other summaries of the model's quality of fit
- (from `augment`) fitted values, residuals and other summaries related to individual points used to fit the model, or individual predictions made by the model, which will be helpful for assessing predictive accuracy and for developing diagnostic tools for assessing the assumptions of multiple regression.

26.2 What is important in 431?

In 431, my primary goal is to immerse you in several cases, which will demonstrate good statistical practice in the analysis of data using multiple regression models. Often, we will leave gaps for 432, but the principal goal is to get you to the point where you can do a solid (if not quite complete) analysis of data for the modeling part (Study 2) of Project B.

Key topics regarding multiple regression we cover in 431 include:

1. Describing the multivariate relationship - Scatterplots and smoothing - Correlation coefficients, Correlation matrices
2. Transformations and Re-expression - The need for transformation - Using a Box-Cox method to help identify effective transformation choices - Measuring and addressing collinearity
3. Testing the significance of a multiple regression model - T tests for individual predictors as last predictor in - Global F tests based on ANOVA to assess overall predictive significance - Incremental and Sequential testing of groups of predictors
4. Interpreting the predictive value of a model - R^2 and Adjusted R^2 , along with AIC and BIC - Residual standard deviation and RMSE
5. Checking model assumptions - Residual Analysis including studentized residuals, and the major plots - Identifying points with high Leverage - Assessing Influence numerically and graphically
6. Model Selection - The importance of parsimony - Stepwise regression
7. Assessing Predictive Accuracy through Cross-Validation - Summaries of predictive error - Plotting predictions across multiple models
8. Summarizing the Key Findings of the Model, briefly and accurately - Making the distinction between causal findings and associations - The importance of logic, theory and empirical evidence. (LTE)

A Getting Data Into R

Using data from an R package

To use data from an R package, for instance, the `bechdel` data from the `fivethirtyeight` package, you can simply load the relevant package with `library` and then the data frame will be available

```
library(fivethirtyeight)
library(tidyverse)

bechdel

# A tibble: 1,794 x 15
  year imdb     title test  clean~1 binary budget domgr~2 intgr~3 code  budge~4
  <int> <chr>   <chr> <chr> <ord>   <chr>  <int>  <dbl> <dbl> <chr>  <int>
1 2013 tt1711~ 21 &~ nota~ notalk FAIL    1.3 e7  2.57e7  4.22e7 2013~  1.3 e7
2 2012 tt1343~ Dred~ ok-d~ ok      PASS    4.5 e7  1.34e7  4.09e7 2012~  4.57e7
3 2013 tt2024~ 12 Y~ nota~ notalk FAIL    2   e7  5.31e7  1.59e8 2013~  2   e7
4 2013 tt1272~ 2 Gu~ nota~ notalk FAIL    6.1 e7  7.56e7  1.32e8 2013~  6.1 e7
5 2013 tt0453~ 42 men   men     FAIL    4   e7  9.50e7  9.50e7 2013~  4   e7
6 2013 tt1335~ 47 R~ men   men     FAIL    2.25e8 3.84e7  1.46e8 2013~  2.25e8
7 2013 tt1606~ A Go~ nota~ notalk FAIL    9.2 e7  6.73e7  3.04e8 2013~  9.2 e7
8 2013 tt2194~ Abou~ ok-d~ ok      PASS    1.2 e7  1.53e7  8.73e7 2013~  1.2 e7
9 2013 tt1814~ Admi~ ok   ok      PASS    1.3 e7  1.80e7  1.80e7 2013~  1.3 e7
10 2013 tt1815~ Afte~ nota~ notalk FAIL   1.3 e8  6.05e7  2.44e8 2013~  1.3 e8
# ... with 1,784 more rows, 4 more variables: domgross_2013 <dbl>,
#   intgross_2013 <dbl>, period_code <int>, decade_code <int>, and abbreviated
#   variable names 1: clean_test, 2: domgross, 3: intgross, 4: budget_2013
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Using `read_rds` to read in an R data set

We have provided the `nnyfs.Rds` data file on the course data page.

Suppose you have downloaded this data file into a directory on your computer called `data` which is a sub-directory of the directory where you plan to do your work, perhaps called `431-nnyfs`.

Open RStudio and create a new project into the `431-nnyfs` directory on your computer. You should see a `data` subdirectory in the Files window in RStudio after the project is created.

Now, read in the `nnyfs.Rds` file to a new tibble in R called `nnyfs_new` with the following command:

```
nnyfs_new <- read_rds("data/nnyfs.Rds")
```

Here are the results...

```
nnyfs_new
```

```
# A tibble: 1,518 x 45
  SEQN sex    age_ch~1 race_~2 educ_~3 langu~4 sampl~5 incom~6 age_a~7 educ_~8
  <dbl> <fct>   <dbl> <fct>   <dbl> <fct>   <dbl> <dbl> <dbl> <dbl> <dbl>
  1 71917 Female     15 3_Blac~      9 English  28299.   0.21    46 2_9-11~
  2 71918 Female     8 3_Blac~      2 English  15127.    5       46 3_High~
  3 71919 Female     14 2_Whit~     8 English  29977.    5       42 5_Coll~
  4 71920 Female     15 2_Whit~     8 English  80652.   0.87    53 3_High~
  5 71921 Male       3 2_Whit~     NA English  55592.   4.34    31 3_High~
  6 71922 Male       12 1_Hisp~     6 English  27365.    5       42 4_Some~
  7 71923 Male       12 2_Whit~     5 English  86673.    5       39 2_9-11~
  8 71924 Female     8 4_Othe~     2 English  39549.   2.74    31 3_High~
  9 71925 Male       7 1_Hisp~     0 English  42333.   0.46    45 2_9-11~
 10 71926 Male      8 3_Blac~     2 English  15307.   1.57    56 3_High~
# ... with 1,508 more rows, 35 more variables: respondent <fct>,
#   salt_used <fct>, energy <dbl>, protein <dbl>, sugar <dbl>, fat <dbl>,
#   diet_yesterday <fct>, water <dbl>, plank_time <dbl>, height <dbl>,
#   weight <dbl>, bmi <dbl>, bmi_cat <fct>, arm_length <dbl>, waist <dbl>,
#   arm_circ <dbl>, calf_circ <dbl>, calf_skinfold <dbl>,
#   triceps_skinfold <dbl>, subscapular_skinfold <dbl>, active_days <dbl>,
#   tv_hours <dbl>, computer_hours <dbl>, physical_last_week <fct>, ...
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Using `read_csv` to read in a comma-separated version of a data file

We have provided the `nnyfs.csv` data file on the course data page.

Suppose you have downloaded this data file into a directory on your computer called `data` which is a sub-directory of the directory where you plan to do your work, perhaps called `431-nnyfs`.

Open RStudio and create a new project into the `431-nnyfs` directory on your computer. You should see a `data` subdirectory in the Files window in RStudio after the project is created.

Now, read in the `nnyfs.csv` file to a new tibble in R called `nnyfs_new2` with the following command:

```
nnyfs_new2 <- read_csv("data/nnyfs.csv")  
  
Rows: 1518 Columns: 45  
-- Column specification -----  
Delimiter: ","  
chr (18): sex, race_eth, language, educ_adult, respondent, salt_used, diet_y...  
dbl (27): SEQN, age_child, educ_child, sampling_wt, income_pov, age_adult, e...  
  
i Use `spec()` to retrieve the full column specification for this data.  
i Specify the column types or set `show_col_types = FALSE` to quiet this message.  
  
nnyfs_new2  
  
# A tibble: 1,518 x 45  
  SEQN sex    age_ch~1 race_~2 educ_~3 langu~4 sampl~5 incom~6 age_a~7 educ_~8  
  <dbl> <chr>   <dbl> <chr>    <dbl> <chr>    <dbl> <dbl>    <dbl> <chr>  
1 71917 Female     15 3_Blac~      9 English  28299.   0.21    46 2_9-11~  
2 71918 Female     8 3_Blac~      2 English  15127.    5       46 3_High~  
3 71919 Female     14 2_Whit~     8 English  29977.    5       42 5_Coll~  
4 71920 Female     15 2_Whit~     8 English  80652.   0.87    53 3_High~  
5 71921 Male       3 2_Whit~     NA English  55592.   4.34    31 3_High~  
6 71922 Male       12 1_Hisp~     6 English  27365.    5       42 4_Some~  
7 71923 Male       12 2_Whit~     5 English  86673.    5       39 2_9-11~  
8 71924 Female     8 4_Othe~     2 English  39549.   2.74    31 3_High~  
9 71925 Male       7 1_Hisp~     0 English  42333.   0.46    45 2_9-11~  
10 71926 Male      8 3_Blac~     2 English  15307.   1.57    56 3_High~  
# ... with 1,508 more rows, 35 more variables: respondent <chr>,  
#   salt_used <chr>, energy <dbl>, protein <dbl>, sugar <dbl>, fat <dbl>,  
#   diet_yesterday <chr>, water <dbl>, plank_time <dbl>, height <dbl>,  
#   weight <dbl>, bmi <dbl>, bmi_cat <chr>, arm_length <dbl>, waist <dbl>,  
#   arm_circ <dbl>, calf_circ <dbl>, calf_s Skinfold <dbl>,
```

```
# triceps_skinfold <dbl>, subscapular_skinfold <dbl>, active_days <dbl>,
# tv_hours <dbl>, computer_hours <dbl>, physical_last_week <chr>, ...
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

If you also want to convert the `character` variables to `factors`, as you will often want to do before analyzing the results, you should instead use:

```
nnyfs_new3 <- read_csv("data/nnyfs.csv") %>%
  mutate(across(where(is.character), as_factor))
```

```
Rows: 1518 Columns: 45
-- Column specification -----
Delimiter: ","
chr (18): sex, race_eth, language, educ_adult, respondent, salt_used, diet_y...
dbl (27): SEQN, age_child, educ_child, sampling_wt, income_pov, age_adult, e...
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
nnyfs_new3
```

```
# A tibble: 1,518 x 45
  SEQN sex    age_ch~1 race_~2 educ_~3 langu~4 sampl~5 incom~6 age_a~7 educ_~8
  <dbl> <fct>   <dbl> <fct>   <dbl> <fct>   <dbl> <dbl>   <dbl> <fct>
  1 71917 Female     15 3_Blac~      9 English  28299.   0.21    46 2_9-11~
  2 71918 Female     8 3_Blac~      2 English  15127.    5       46 3_High~
  3 71919 Female     14 2_Whit~     8 English  29977.    5       42 5_Coll~
  4 71920 Female     15 2_Whit~     8 English  80652.   0.87    53 3_High~
  5 71921 Male       3 2_Whit~     NA English  55592.   4.34    31 3_High~
  6 71922 Male       12 1_Hisp~     6 English  27365.    5       42 4_Some~
  7 71923 Male       12 2_Whit~     5 English  86673.    5       39 2_9-11~
  8 71924 Female     8 4_Othe~     2 English  39549.   2.74    31 3_High~
  9 71925 Male       7 1_Hisp~     0 English  42333.   0.46    45 2_9-11~
 10 71926 Male      8 3_Blac~     2 English  15307.   1.57    56 3_High~
# ... with 1,508 more rows, 35 more variables: respondent <fct>,
# salt_used <fct>, energy <dbl>, protein <dbl>, sugar <dbl>, fat <dbl>,
# diet_yesterday <fct>, water <dbl>, plank_time <dbl>, height <dbl>,
# weight <dbl>, bmi <dbl>, bmi_cat <fct>, arm_length <dbl>, waist <dbl>,
# arm_circ <dbl>, calf_circ <dbl>, calf_skinfold <dbl>,
# triceps_skinfold <dbl>, subscapular_skinfold <dbl>, active_days <dbl>,
```

```
#   tv_hours <dbl>, computer_hours <dbl>, physical_last_week <fct>, ...
# i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Note that, for example, `sex` and `race_eth` are now listed as factor (`fctr`) variables. One place where this distinction between `character` and `factor` variables matters is when you summarize the data.

```
summary(nnyfs_new2$race_eth)
```

Length	Class	Mode
1518	character	character

```
summary(nnyfs_new3$race_eth)
```

3_Black Non-Hispanic	2_White Non-Hispanic	1_Hispanic
338	610	450
4_Other Race/Ethnicity		
120		

Converting Character Variables into Factors

The command you want to create `newdata` from `olddata` is:

```
newdata <- olddata %>%
  mutate(across(where(is.character), as_factor))
```

For more on factors, visit <https://r4ds.had.co.nz/factors.html>

Converting Data Frames to Tibbles

Use `as_tibble()` or simply `tibble()` to assign the attributes of a tibble to a data frame. Note that `read_rds` and `read_csv` automatically create tibbles.

For more on tibbles, visit <https://r4ds.had.co.nz/tibbles.html>.

For more advice

Consider visiting the software tutorials page under the R and Data heading on our main web site.

B References

- Baumer, Benjamin S., Daniel T. Kaplan, and Nicholas J. Horton. 2017. *Modern Data Science with r*. Boca Raton, FL: CRC Press. <https://mdsr-book.github.io/>.
- Bernard, Gordon R., Arthur P. Wheeler, James A. Russell, Roland Schein, Warren R. Summer, Kenneth P. Steinberg, William J. Fulkerson, et al. 1997. “The Effects of Ibuprofen on the Physiology and Survival of Patients with Sepsis.” *New England Journal of Medicine* 336: 912–18. <http://www.nejm.org/doi/full/10.1056/NEJM199703273361303#t=article>.
- Bock, David E., Paul F. Velleman, and Richard D. De Veaux. 2004. *Stats: Modelling the World*. Boston MA: Pearson Addison-Wesley.
- Dupont, William D. 2002. *Statistical Modeling for Biomedical Researchers*. New York: Cambridge University Press.
- Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel-Hierarchical Models*. New York: Cambridge University Press. <http://www.stat.columbia.edu/~gelman/arm/>.
- Gelman, Andrew, and Deborah Nolan. 2017. *Teaching Statistics: A Bag of Tricks*. Second Edition. Oxford, UK: Oxford University Press.
- Good, Phillip I. 2005. *Introduction to Statistics Through Resampling Methods and r/s-PLUS*. Hoboken, NJ: Wiley.
- Ismay, Chester, and Albert Y. Kim. 2022. *ModernDive: Statistical Inference via Data Science*. <http://moderndive.com/>.
- Morton, D., A. Saah, S. Silberg, W. Owens, M. Roberts, and M. Saah. 1982. “Lead Absorption in Children of Employees in a Lead Related Industry.” *American Journal of Epidemiology* 115: 549–55.
- Norman, Geoffrey R., and David L. Streiner. 2014. *Biostatistics: The Bare Essentials*. Fourth. People’s Medical Publishing House.
- Pagano, Marcello, and Kimberlee Gauvreau. 2000. *Principles of Biostatistics*. Second. Duxbury Press.
- Pruzek, Robert M., and James E. Helmreich. 2009. “Enhancing Dependent Sample Analyses with Graphics.” *Journal of Statistics Education* 17(1). <http://ww2.amstat.org/publications/jse/v17n1/helmreich.html>.
- Ramsey, Fred L., and Daniel W. Schafer. 2002. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Second. Pacific Grove, CA: Duxbury.
- Vittinghoff, Eric, David V. Glidden, Stephen C. Shiboski, and Charles E. McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Second. Springer-Verlag, Inc. <http://www.biostat.ucsf.edu/vgsm/>.
- Wainer, Howard. 1997. *Visual Revelations: Graphical Tales of Fate and Deception from*

- Napoleon Bonaparte to Ross Perot.* New York: Springer-Verlag.
- . 2005. *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. Princeton, NJ: Princeton University Press.
- . 2013. *Medical Illuminations: Using Evidence, Visualization and Statistical Thinking to Improve Healthcare*. New York: Oxford University Press.
- Wickham, Hadley, and Garrett Grolemund. 2022. *R for Data Science*. Second. O'Reilly. <https://r4ds.hadley.nz/>.
- Yamada, SB, and EG Boulding. 1998. "Claw Morphology, Prey Size Selection and Foraging Efficiency in Generalist and Specialist Shell-Breaking Crabs." *Journal of Experimental Marine Biology and Ecology* 220: 191–211. http://www.science.oregonstate.edu/~yamadas/SylviaCV/BehrensYamada_Boulding1998.pdf.