

# Thinking with Data (Fourth Edition)

---

John R. Vokey and Scott W. Allen  
Department of Psychology  
The University of Lethbridge

Thinking With Data  
(Fourth Edition)  
Copyright © 1999–2007

John R. Vokey and Scott W. Allen  
A PsyPro book  
Psyence Ink: Lethbridge, Alberta  
(Fall, 2007 printing)

In memory of  
Jean A. M. Simpson Vokey,

a bibliophile and text addict,  
who inhaled words and  
revelled in the worlds thus created



# Preface

This book comprises a collection of lecture notes for the statistics component of the course *Psychology 2030: Methods and Statistics* from the Department of Psychology at the University of Lethbridge. In addition to basic statistical methods, the book includes discussion of many other useful topics. For example, it has a section on writing in APA format (see Chapter 17), and another on how to read the professional psychological literature (see Chapter 16). We even provide a subsection on the secret to living to be 100 years of age (see section A.2.2)—although the solution may not be fully satisfactory!

Despite this volume comprising the fourth edition of the book, it is still very much a work in progress, and is by no means complete. However, despite its current limitations, we expect that students will find it to be a useful adjunct to the lectures. We welcome any suggestions on additions and improvements to the book, and, of course, the report of any typos and other errors.<sup>1</sup> Please email any such errors or corrections to: [vokey@uleth.ca](mailto:vokey@uleth.ca) or [allens@uleth.ca](mailto:allens@uleth.ca).

The book is produced at the cost of printing and distribution, and will evolve and change from semester to semester, so it will have little or no resale value; the latest version is also always available on the web as a portable document format (pdf) file at: <http://people.uleth.ca/~vokey/pdf/thinking.pdf>.

We expect the book to be used, or better, *used up* over the course. In the hands of the student, this volume is intended to become the private statistics diary of the student/reader. To that end, we have printed the book with especially wide, out-side margins, intended to be the repository by the student/reader of further notes and explanatory detail obtained from lecture, private musings, and, as beer is inextricably and ineluctably tied to statistics, pub conversation over a pint or two.

John R. Vokey and Scott W. Allen  
June 28, 2007

---

<sup>1</sup>For example, the Fall, 2005 printing of the 4th edition corrected a series of typos and errors that previous students kindly pointed out to us. This Fall, 2007 printing contains even more corrections, again reported by our sharp-eyed students. Thank you to all!



# Contents

<b>Preface</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Sir Carl Friedrich Gauss (1777-1855)	1
1.2 Summing any Constant Series	3
1.2.1 Definition of a constant series	3
1.2.2 What about constant series with $c > 1$ ?	3
1.3 What's the Point?	4
1.4 Questions	5
<b>I Description</b>	<b>7</b>
<b>2 The Mean (and related statistics)</b>	<b>9</b>
2.1 Definition	9
2.1.1 The mean is a statistic	10
2.2 Properties	10
2.2.1 The mean is that point from which the sum of deviations is zero.	10
2.2.2 The mean as a balancing-point	12
2.2.3 The mean is the point from which the sum of <i>squared</i> deviations is a minimum.	12
2.2.4 And the Mean is ...	13
2.2.5 The Method of Provisional Means	15
2.3 Other Means	16
2.3.1 The Absolute Mean	16
2.3.2 Root-Mean-Square	17
2.3.3 Geometric Mean	17
2.3.4 Harmonic Mean	18
2.4 The Median	19
2.4.1 Definition of the Median	19
2.4.2 A complication	20
2.4.3 Properties of the Median	21
2.5 Questions	21

<b>3</b>	<b>Measures of Variability</b>	<b>25</b>
3.1	The Range . . . . .	25
3.2	The $D^2$ and $D$ statistics . . . . .	26
3.3	Variance ( $S^2$ ) and Standard Deviation ( $S$ ) . . . . .	26
3.4	Questions . . . . .	33
<b>4</b>	<b>Transformed Scores</b>	<b>35</b>
4.1	The Linear Transform . . . . .	35
4.1.1	Rules for changing $\bar{X}$ , $S_X^2$ and $S_X$ . . . . .	35
	Adding a constant to every score . . . . .	36
	Multiplying each score by a constant . . . . .	37
4.2	The Standard Score Transform . . . . .	38
4.2.1	Properties of Standard Scores . . . . .	38
4.2.2	Uses of Standard Scores . . . . .	39
4.3	Questions . . . . .	40
<b>5</b>	<b>Other Descriptive Statistics</b>	<b>43</b>
5.1	Other Statistics . . . . .	43
5.1.1	Skewness . . . . .	43
5.1.2	Kurtosis . . . . .	44
5.2	Questions . . . . .	45
<b>6</b>	<b>Recovering the Distribution</b>	<b>47</b>
6.1	The Markov Inequality . . . . .	47
6.2	The Tchebycheff Inequality . . . . .	50
6.3	Questions . . . . .	51
<b>7</b>	<b>Correlation</b>	<b>53</b>
7.1	Pearson product-moment correlation coefficient . . . . .	53
7.1.1	Sums of Cross-Products . . . . .	54
7.1.2	Sums of Differences . . . . .	56
	What values can ASD take? . . . . .	57
7.2	Covariance . . . . .	58
7.3	Other Correlational Techniques . . . . .	58
7.3.1	The Spearman Rank-Order Correlation Coefficient . . . . .	60
	Tied Ranks . . . . .	61
7.3.2	The Point-Biserial Correlation Coefficient . . . . .	61
7.3.3	And Yet Other Correlational Techniques . . . . .	62
7.4	Questions . . . . .	63
<b>8</b>	<b>Linear Regression</b>	<b>65</b>
8.1	Linear Regression . . . . .	65
8.1.1	The Regression Equation: $Z'_{Y_i} = r_{xy}Z_{X_i}$ . . . . .	67
8.1.2	From standard scores to raw scores . . . . .	68
8.1.3	Correlation and Regression: $r_{xy} = r_{y'y}$ . . . . .	69

8.1.4	The standard error of estimate . . . . .	69
8.1.5	The Proportional Increase in Prediction: PIP . . . . .	73
8.2	Questions . . . . .	74
<b>9</b>	<b>Partial and Multiple Correlation</b>	<b>79</b>
9.1	Partial Correlation . . . . .	79
9.1.1	Part or Partial correlation . . . . .	79
9.1.2	Semi-partial correlation . . . . .	80
9.1.3	An example of Partial Correlation . . . . .	81
9.2	Multiple Correlation . . . . .	82
9.2.1	The Correlation Matrix . . . . .	82
9.2.2	Euler diagrams or “Ballantine’s” . . . . .	83
9.2.3	Computing Multiple-R . . . . .	86
9.3	Questions . . . . .	87
<b>II</b>	<b>Significance</b>	<b>89</b>
<b>10</b>	<b>Introduction to Significance</b>	<b>91</b>
10.1	The Canonical Example . . . . .	94
10.1.1	Counting Events . . . . .	94
	The fundamental counting rule . . . . .	94
	Permutations . . . . .	95
	Combinations . . . . .	95
10.1.2	The Example . . . . .	96
10.1.3	Permutation/Exchangeability Tests . . . . .	97
10.2	Questions . . . . .	98
<b>11</b>	<b>The Binomial Distribution</b>	<b>101</b>
11.0.1	Type I and Type II errors . . . . .	103
	One and two-tailed tests . . . . .	103
	Type I errors and $\alpha$ . . . . .	104
	Type II errors and $\beta$ . . . . .	105
11.0.2	Null and Alternative Hypotheses, and Power . . . . .	105
11.0.3	The Sign Test . . . . .	107
11.1	Questions . . . . .	107
<b>12</b>	<b>The Central Limit Theorem</b>	<b>111</b>
12.1	The Normal Distribution . . . . .	111
12.1.1	The Shah (1985) approximation . . . . .	118
12.2	The Normal Approximation to the Binomial . . . . .	118
12.3	Sums From Any Distributions . . . . .	120
12.3.1	Sums of Normal Distributions . . . . .	120
12.3.2	The Standard Error of the Mean . . . . .	121
12.3.3	Sums of Nonnormal Distributions . . . . .	121
12.3.4	A Confidence Trick . . . . .	122

12.3.5	Why it's called the "Normal Distribution" . . . . .	123
12.4	Questions . . . . .	125
<b>13</b>	<b>The <math>\chi^2</math> Distribution</b>	<b>127</b>
13.1	Chi square and the goodness of fit . . . . .	127
13.2	$\chi^2$ tests of independence . . . . .	132
13.2.1	2 x 2 contingency tables . . . . .	132
	Degrees of freedom . . . . .	133
	A measure of association . . . . .	134
13.2.2	r x c contingency tables . . . . .	134
	A measure of association for r x c tables . . . . .	135
13.3	Questions . . . . .	135
<b>14</b>	<b>The <i>t</i> Distribution</b>	<b>139</b>
14.1	Student's <i>t</i> -test . . . . .	139
14.1.1	<i>t</i> -test for one sample mean . . . . .	143
14.1.2	<i>t</i> -test for two <i>dependent</i> sample means . . . . .	145
14.2	<i>t</i> -test for two <i>independent</i> sample means . . . . .	146
14.2.1	Relationship between <i>t</i> and $r_{pb}$ . . . . .	148
14.3	Questions . . . . .	148
<b>15</b>	<b>The General Logic of ANOVA</b>	<b>151</b>
15.1	An example . . . . .	151
15.1.1	Fundamental ANOVA equation . . . . .	152
15.1.2	Mean squares . . . . .	153
15.1.3	The <i>F</i> -ratio . . . . .	155
15.2	<i>F</i> and <i>t</i> . . . . .	156
15.3	Questions . . . . .	156
<b>III</b>	<b>The Psychological Literature</b>	<b>159</b>
<b>16</b>	<b>The Literature</b>	<b>161</b>
16.1	Meetings . . . . .	161
16.1.1	Talks . . . . .	162
16.1.2	Posters . . . . .	163
16.2	Journals . . . . .	163
16.3	Books . . . . .	164
<b>17</b>	<b>The APA Format</b>	<b>165</b>
17.0.1	Title . . . . .	166
17.0.2	Abstract . . . . .	166
17.0.3	Introduction . . . . .	167
17.0.4	Method . . . . .	167
	Participants . . . . .	167
	Materials . . . . .	168

Procedure . . . . .	168
Details, details . . . . .	168
17.0.5 Results . . . . .	168
17.0.6 Discussion . . . . .	169
17.0.7 References . . . . .	169
17.0.8 Notes . . . . .	170
17.0.9 Tables . . . . .	171
17.0.10 Figures . . . . .	171
17.0.11 Appendices . . . . .	171
<b>IV Appendices</b>	<b>173</b>
<b>A Summation</b>	<b>175</b>
A.1 What is Summation? . . . . .	175
A.1.1 The Summation Function . . . . .	175
A.1.2 The Summation Operator: $\sum$ . . . . .	176
A.2 Summation Properties . . . . .	176
A.2.1 Summation is commutative . . . . .	177
A.2.2 Summation is associative . . . . .	177
The secret to living to 100 years of age . . . . .	178
A.3 Summation Rules . . . . .	178
A.3.1 Rule 1: $\sum(X + Y) = \sum X + \sum Y$ . . . . .	178
A.3.2 Rule 2: $\sum(X - Y) = \sum X - \sum Y$ . . . . .	179
A.3.3 Rule 3: $\sum(X + c) = \sum X + nc$ . . . . .	179
A.3.4 Rule 4: $\sum_{i=1}^n c = nc$ . . . . .	180
A.3.5 Rule 5: $\sum cX = c \sum X$ . . . . .	180
A.4 Questions . . . . .	180
<b>B Keppel's System for Complex ANOVAs</b>	<b>181</b>
B.1 Identifying Sources of Variation . . . . .	182
B.2 Specifying Degrees of Freedom . . . . .	182
B.3 Generating Expected Mean Squares . . . . .	183
B.4 Selecting Error-Terms . . . . .	184
<b>C Answers to Selected Questions</b>	<b>185</b>
C.1 Chapter 1: Page 5 . . . . .	185
C.2 Chapter 2: Page 21 . . . . .	186
C.3 Chapter 3: Page 33 . . . . .	186
C.4 Chapter 4: Page 40 . . . . .	186
C.5 Chapter 5: Page 45 . . . . .	187
C.6 Chapter 6: Page 51 . . . . .	187
C.7 Chapter 7: Page 63 . . . . .	187
C.8 Chapter 8: Page 74 . . . . .	187
C.9 Chapter 9: Page 87 . . . . .	187

C.10 Chapter 10: Page 98 . . . . .	188
C.11 Chapter 11: Page 107 . . . . .	188
C.12 Chapter 12: Page 125 . . . . .	188
C.13 Chapter 13: Page 135 . . . . .	189
C.14 Chapter 14: Page 148 . . . . .	189
C.15 Chapter 15: Page 156 . . . . .	189
C.16 Appendix A: Page 180 . . . . .	190
<b>References</b>	<b>191</b>
<b>Index</b>	<b>192</b>

# List of Figures

1.1	Sir Carl Friedrich Gauss . . . . .	2
2.1	The mean is that point from which the sum of deviations is zero	11
2.2	The mean minimises the sum of squares . . . . .	13
3.1	A map of distances between towns and cities along the Trans-Canada highway. . . . .	27
6.1	Andrei Andreyevich Markov . . . . .	48
6.2	Pafnuti L. Tchebycheff . . . . .	50
7.1	Karl Pearson . . . . .	54
8.1	Sir Francis Galton . . . . .	66
8.2	Scatterplot and regression lines for the hypothetical data from Table 8.1. . . . .	70
8.3	Scatterplot and regression lines for the standardized hypothetical data from Table 8.1. . . . .	71
8.4	Proportional increase in prediction as a function of $ r_{xy} $ . . . . .	75
9.1	The label from one of the brands of beer produced by P. Ballantine & Sons depicting the overlapping circle logo. . . . .	84
9.2	A Euler diagram or “Ballantine” of the example correlations from Table 9.2. The area of overlap of any one circle with another represents the squared correlation between the two variables. . .	85
9.3	The same figure as in Figure 9.2, except shaded to depict the area of interest. . . . .	85
10.1	Sir Ronald Aylmer Fisher . . . . .	92
10.2	The frequency distribution of the 70 mean differences between the male and female sets of scores produced from the possible 70 unique exchanges of male and female scores. . . . .	99
11.1	Relative frequency as a function of the statistic $Z$ . A plot of the binomial distribution from Table 11.1. . . . .	104

11.2	Examples of binomial distributions . . . . .	106
12.1	The standard normal curve with $\mu = 0$ and $\sigma = 1$ . . . . .	112
12.2	The normal approximation to the binomial. . . . .	119
13.1	$\chi^2$ distribution . . . . .	129
14.1	W. S. Gosset (“Student”) . . . . .	140
15.1	Plot of the $F$ -distribution with 2 and 6 degrees of freedom. . . . .	155

# List of Tables

2.1	Demonstration using summation notation that the mean is that point from which the sum of deviations is zero. . . . .	11
2.2	Demonstration that the sum of squared deviations from the mean is a minimum. . . . .	14
3.1	Distances along the Trans-Canada Highway . . . . .	28
3.2	Distances from Calgary . . . . .	29
3.3	Derivation of the computational formula for variance. . . . .	31
3.4	Demonstration of the equivalence of $D^2 = 2S^2$ . . . . .	32
7.1	Demonstration that the average squared difference formula for $r$ and the cross-products formula for $r$ are equivalent. . . . .	59
7.2	Demonstration that the Spearman formula is equivalent to the Pearson cross-product formula applied to the data in ranks (equation 7.1). . . . .	60
7.3	Demonstration that the point-biserial formula is equivalent to applying the Pearson cross-product formula to the data for which the categorical variable is given a binary code (equation 7.1). . . . .	62
8.1	Example regression data. . . . .	69
8.2	Proportional increase in prediction: PIP. . . . .	74
9.1	Derivation of the expression for partial correlation. . . . .	80
9.2	The correlation matrix of the example variables of smoking (cigarette sales), year, and cancer (lung-cancer death rate 20 years later). . . . .	82
10.1	A sampling of the 70 possible exchanges of the male and female scores. . . . .	98
11.1	The binomial distribution. . . . .	103
12.1	Areas $[\Phi(z)]$ under the Normal Curve . . . . .	113
13.1	$\chi^2$ Table . . . . .	130

13.2	The observed and expected (in parentheses) frequencies of national political party preferences for the example polynomial situation discussed in the text. . . . .	131
13.3	An example of a 2 x 2 contingency table depicting the cross-classification of cola preference by sex. . . . .	132
13.4	An example of a 2 x 2 contingency table in which the cross-classifications of cola preference and sex are independent. . . . .	133
14.1	An example of bias in using sample variances ( $S^2$ ) as estimates of population variance. . . . .	141
14.2	Critical values of the $t$ -distribution. . . . .	144
15.1	Hypothetical values for three observations per each of three groups.	152
15.2	Sums of squares for the hypothetical data from Table 15.1. . . . .	153
15.3	Critical values of the $F$ -distribution . . . . .	157
B.1	Hypothetical research design. . . . .	182
B.2	Calculation of the degrees of freedom for the research design in Table B.1. . . . .	183
B.3	The expected mean squares for the sources of variance in the design shown in Table B.1. . . . .	184

# Chapter 1

## Introduction

### 1.1 Sir Carl Friedrich Gauss (1777-1855)

As an old man, the eminent mathematician, Sir Carl Friedrich Gauss (see Figure 1.1) enjoyed relating the story of his first day in his first class in arithmetic at the age of 10. His teacher, Herr Büttner, was a brutish man who was a firm believer in the pedagogical advantages of thrashing the boys under his tutelage. Having been their teacher for many years, Büttner knew that none of the boys had ever heard about arithmetic progressions or series. Accordingly, and in the fashion of bullies everywhere and everywhen who press every temporary advantage to prove their superiority, he set the boys a very long problem of addition that he, unbeknownst to boys, could answer in seconds with a simple formula.

The problem was of the sort of adding up the numbers  $176 + 195 + 214 + \dots + 2057$ , for which the step from each number to the next is the same (here, 19), and a fixed number of such numbers (here 100) are to be added. As was the tradition of the time, the students were instructed to place their slates<sup>1</sup> on a table, one slate on top of the other in order, as soon as each had each completed the task. No sooner had Büttner finished the instructions than Gauss placed his slate on the table, saying “There it lies.”<sup>2</sup> Incredulous (or so it goes), the teacher looked at him scornfully (but, no doubt, somewhat gleefully as he anticipated the beating he would deliver to Gauss for what surely must be a wrong answer) while the other students continued to work diligently for another hour. Much later, when the teacher finally checked the slates, Gauss was the only student to have the correct answer (Bell, 1937, pp. 221–222).<sup>3</sup>

---

<sup>1</sup>Yes, in the late 18th century, slates, not paper were used by schoolchildren.

<sup>2</sup>Actually, as the son of a poor, German family, he no doubt said it in his peasant German dialect: “Ligget se”.

<sup>3</sup>Büttner was so impressed with the obvious brilliance of Gauss that at least for Gauss he became a humane and caring teacher. Out of his own limited resources, Büttner purchased the best book on arithmetic of the day and gave it to Gauss. He also introduced the 10-year



Figure 1.1: Sir Carl Friedrich Gauss

How did he do it? For ease of exposition, let us assume the simpler task of summing the integers from 1 to 100 inclusive. As the story goes, Gauss simply imagined the sum he sought, say  $S$ , as being written simultaneously in both ascending and descending order:

$$S = 1 + 2 + 3 + \cdots + 98 + 99 + 100$$

$$S = 100 + 99 + 98 + \cdots + 3 + 2 + 1$$

Then, instead of adding the numbers horizontally across the rows, he added them vertically:

$$S + S = (1 + 100) + (2 + 99) + (3 + 98) + \cdots + (98 + 3) + (99 + 2) + (100 + 1)$$

or

$$S + S = 101 + 101 + 101 + \cdots + 101 + 101 + 101$$

Because there are 100 sums of 101, twice the desired sum must be  $100 * 101$  or 10100:

$$2S = 100 * 101 = 10100$$

---

old Gauss to his young (age 17), mathematically-inclined assistant, Johann Martin Bartels (1769–1836) to study with Gauss when it was clear that Gauss' skills were beyond those of Büttner. Gauss and Bartels remained life-long friends.

And if  $2S = 10100$ , then  $S$  must equal  $\frac{10100}{2} = 5050!$

## 1.2 Summing any Constant Series

We can generalise this example into a principle for summing *any* series of numbers.

### 1.2.1 Definition of a constant series

First, we define a series of numbers as any ordered set of numbers,  $X$ , for which

$$X_{i+1} - X_i = c \quad (1.1)$$

for all  $X_i$ . That is, the difference between any number in the set,  $X_i$ , and the next number,  $X_{i+1}$ , is a constant,  $c$ . For example,  $X = \{1, 2, 3, \dots, 6\}$  is a series by this definition with  $c = 1$ . So is  $X = \{1, 3, 5, 7\}$ , except that  $c = 2$ . In general, with the set,  $X$ , defined as in equation 1.1, we can ask: what is the sum of  $X$ ?

Following the 10-year old Gauss, we note that the sum across each of the pairs formed by lining up the ascending series with its descending counterpart is always a constant,  $k$ , for any constant series. In the case of summing the numbers from 1 to 100,  $k = 101$ . As any of the pairs thus formed for a given set yields the same sum,  $k$ , we can define  $k = X_1 + X_n$ , that is, as the sum of the first and last numbers of the set. We also note that in doing the pairing, we have used each number in the pair twice. That is, using  $n$  to indicate the number of numbers in the series, in general,  $nk =$  twice the desired sum. Therefore,

$$X_1 + X_2 + \dots + X_n = n \left( \frac{X_1 + X_n}{2} \right) \quad (1.2)$$

### 1.2.2 What about constant series with $c > 1$ ?

We immediately note a problem in using equation (1.2), introduced with values of  $c$  greater than 1. Although not so obvious with a series such as  $X = \{1, 3, 5, 7\}$ , for which  $n$  obviously equals 4, the problem is much clearer with a series such as  $X = \{4, 7, \dots, 31\}$ , or the example series  $176 + 195 + 214 + \dots + 2057$  presented earlier. In these latter cases, how do we determine the value of  $n$ ; that is, how many numbers are in the series? If we are reduced to counting on our fingers to determine  $n$  for such series, then Gauss's childhood insight is of limited use.

However,  $X_n - X_1$  yields the distance between the last number,  $X_n$ , and the first number,  $X_1$ , in the series. If  $c = 1$ , then this distance equals the number of numbers, less one (e.g.,  $100 - 1 = 99$ , plus 1 equals 100—the number of numbers between 1 and 100, inclusive). If  $c = 2$ , then this distance equals *twice* the number of numbers in the series, less one; so, dividing by two and adding one would yield the correct value of  $n$ . If  $c = 3$ , then this distance equals *three* times the number of items in the series, less one; so, dividing by three and adding one

would yield the correct value of  $n$ . And so on. In general,

$$n = \frac{X_n - X_1}{c} + 1 \quad (1.3)$$

To be completely general, then, Gauss's formula for the sum of a constant series is written as

$$X_1 + X_2 + \dots + X_n = \underbrace{\left( \frac{X_n - X_1}{c} + 1 \right)}_n \underbrace{\left( \frac{X_1 + X_n}{2} \right)}_{\frac{k}{2}} \quad (1.4)$$

For example, applying equation (1.4) to obtain the sum of the series,  $X = \{4, 7, \dots, 31\}$ , we note that  $c = 3$  and that, therefore,  $n = \frac{31-4}{3} + 1 = 10$ . As  $k = X_1 + X_n = 31 + 4 = 35$ , and  $\frac{k}{2} = \frac{35}{2} = 17.5$ , then the sum equals  $10(17.5) = 175$ . For the sum of the earlier series,  $176 + 195 + 214 + \dots + 2057$ ,  $c = 195 - 176 = 19$ , so  $n = \frac{2057-176}{19} + 1 = 100$ ,  $k = 2057 + 176 = 2233$ , so the sum is  $100 \frac{2233}{2} = 111650$ .

### 1.3 What's the Point?

We relate the story of Gauss's childhood insight to underscore a central theme of the exposition of the statistical material in this book. Arithmetic and mathematics more generally often reveal surprising properties of numbers that, with a little insight, turn out to be quite useful for dealing with sets of numbers we encounter as data. As applied statistics is first and foremost a mathematical discipline that concerns itself with characterising sets of numbers, it consists in the main of the exploitation of these useful arithmetical properties in the form of statistical formulae and equations. Rather than presenting these formulae as if given by divine revelation or as magic black boxes to be fed raw data in return for the output of "statistics" (both of which we have found far too often to be the case in introductory textbooks on statistics), where possible (without stretching too far afield), we emphasise the arithmetic property or properties being exploited, so that despite any superficial complexity, the underlying simplicity and insight of the formula or technique may be appreciated.

This approach means that the style of exposition moves in a direction contrary to that of many recent textbooks on the same topic. Rather than suppress the underlying mathematics, as many of them do, we revel in it, with pages of derivations and walked-through "proofs". To make this approach work, and yet still maintain an introductory level, we restricted all derivations to elementary algebra, eschewing the often more elegant derivations and "proofs" available with differential and integral calculus. So be it. In our view, it is more important to induce an understanding and appreciation than to reach for an elegance that, for many, only obfuscates. To that end, we also provide examples of each of the key computations in the form of precisely focussed questions at the end of each chapter, along with the worked-through solutions for a selection of these

questions (see Appendix C), so that the student may explore each of the concepts in a concrete way.

But the story of Gauss's childhood insight highlights another important point. These useful properties are properties of *numbers*, and, as such apply to numbers—any and all numbers. These properties are *not* properties of what the numbers on any given occasion refer to, or *mean*. Whether it makes sense to apply one of these techniques to a given set of numbers is not a question of mathematics. Gauss's insight provides the sum of any constant series of numbers whether it makes sense (in terms of what the numbers refer to) to take such a sum or not. All of the statistical techniques we discuss are properties of the numbers themselves. Extension of these properties to the referents of the numbers requires an *inference* that is not mathematical in nature. Part of the polemic of this book is to make this distinction and its consequences clear.

## 1.4 Questions

1. What is the arithmetic average of the series  $X = \{11, 19, 27, \dots, 75\}$ ?
2. How many numbers are in the series  $X = \{-0.95, -0.91, -0.87, \dots, 0.73\}$ ?
3. For a series of 19 numbers,  $X$ ,  $X_{i+1} = X_i + .5$ , and  $X_1 = -3$ , what is  $X_{19}$ ?



**Part I**

**Description**



## Chapter 2

# The Mean (and related statistics)

### 2.1 Definition

A *mean* is a *normalised sum*; which is to say that a mean of a set of scores is just the sum of that set of scores corrected for the number of scores in the set. Normalising the sums in this way allows for the direct comparison of one normalised sum with another without having to worry about whether the sums differ (or not) from each other simply because one sum is based on more numbers than the other. One sum, for example, could be larger than another simply because it is based on many more numbers rather than because it was derived from, in general, larger numbers than another sum. Expressed as means, however, any difference between two sums must directly reflect the general or “average” magnitude of the numbers on which they are based.

You probably know the mean as the *arithmetic average*: the sum of the scores divided by the number of scores. In summation notation,<sup>1</sup> the mean can be defined as:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (2.1)$$

As shown in equation 2.1, the symbol for the mean of a set of scores is the name of the set of scores with an overbar; if the name of the set of scores is  $X$ , then the symbol of the mean of that set is  $\bar{X}$ . If, on the other hand, the name of the set were “age” or “height” or  $Z$ , say, then the respective means would be labelled as  $\overline{\text{age}}$ ,  $\overline{\text{height}}$ , and  $\bar{Z}$ .

---

<sup>1</sup>See Appendix A for a review of summation and the summation operator,  $\Sigma$ .

### 2.1.1 The mean is a statistic

As with other simple numerical summaries of the data, such as the *median* discussed subsequently, the mean is a *statistic* used to *locate* the distribution of scores along the number line. A *statistic* is simply any numerical summary statement about a set or distribution of scores. The mean is typically the statistic quoted in statements that begin “In general, . . .” or “On average, . . .” when referring to a set of scores, and indicates in a general way whether the distribution is located among the high numbers or the low numbers, negative or positive.

## 2.2 Properties

Although no doubt often thought of contemptuously given its familiarity,<sup>2</sup> the mean has useful properties not shared with any other single-valued summary of the data.

### 2.2.1 The mean is that point from which the sum of deviations is zero.

This property equally could have been used as the definition of the mean. That is, we could have defined the mean as that value from which the sum of deviations is zero and then derived equation 2.1 from it as, say, the “calculation” formula. Not that we couldn’t use this property as a method to find the mean of a distribution of numbers; it just wouldn’t be very efficient. Still, if you have found a value that just happens to result in the sum of deviations from it being zero, you *know* that that value equals the mean by *definition*.

This property is probably most easily grasped graphically, as shown in Figure 2.1. For any distribution of scores, a plot of the sum of deviation scores as a function of different values or “guesses” about what the mean might be results in a diagonal line running from positive sums on the left for guesses less than the mean to negative sums on the right for guesses greater than the mean. The “guess” corresponding to the point at which the line crosses a sum of zero is the mean of the distribution.

What this property means is that *any other value whatsoever* must result in a sum of deviations from it that is different from zero. Expressed in summation notation,

$$\sum_{i=1}^n (X_i - \bar{X}) = 0 \quad (2.2)$$

This property is demonstrated in terms of the summation rules in Table 2.1.

---

<sup>2</sup>As the saying goes, “Familiarity breeds contempt.”

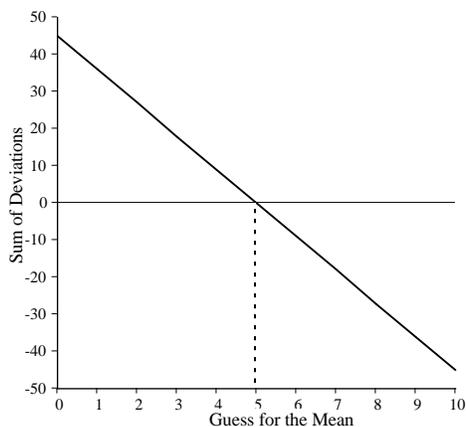


Figure 2.1: The mean is that point from which the sum of deviations is zero. In this example, which plots the sum of deviations as a function of different “guesses” for the mean,  $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$  and  $\bar{X} = 5$ .

First,

$$\sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n X_i - \sum_{i=1}^n \bar{X}$$

As  $\sum_{i=1}^n c = nc$ , then

$$\sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n X_i - n\bar{X}$$

But,  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ , so

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X}) &= \sum_{i=1}^n X_i - n \frac{\sum_{i=1}^n X_i}{n} \\ &= \sum_{i=1}^n X_i - \sum_{i=1}^n X_i \\ &= 0 \end{aligned}$$

Table 2.1: Demonstration using summation notation that the mean is that point from which the sum of deviations is zero.

### 2.2.2 The mean as a balancing-point

This property also means that if we create a new set of scores, called *deviation scores*, by subtracting the mean from each score in the set, the *sum* and therefore the *mean* of the deviation scores are necessarily equal to zero. That is, the mean is the *balancing-point* of a distribution of scores: the sum of the deviations above the mean precisely balances the sum of the deviations below the mean. For this reason, a distribution of scores obtained by subtracting the mean from every score is referred to as a *centred distribution*, and the act of producing such deviation scores as one of *centring* the distribution.

For example, for  $X = \{1, 2, 3, 6, 8\}$ , the mean is  $20/5 = 4$ . Converting each of the scores of  $X$  to a deviation score by subtracting the mean from each of them yields  $x = \{-3, -2, -1, 2, 4\}$ .<sup>3</sup> The sum of the scores below the mean equals the negative of the sum of those above the mean (i.e.,  $-3 + -2 + -1 = -[2 + 4]$ ) and, hence, balances it.

### 2.2.3 The mean is the point from which the sum of *squared* deviations is a minimum.

As with the previous property, this property also could have served as the definition of the mean. That is, we could have used this property to define the mean and then derived both equations 2.1 and 2.2 from it. Similarly, again as with the previous property, this property of the mean is probably most easily understood graphically. If the sum of squared deviations is plotted as a function of different values or guesses for the mean, a parabola or bowl-shaped curve is produced. The guess corresponding to the exact bottom of this bowl, the minimum of the function, would be the mean (see Figure 2.2).

Using  $Y$  as *any value whatsoever*, this property is expressed in summation notation as:

$$\sum_{i=1}^n (X_i - \bar{X})^2 \leq \sum_{i=1}^n (X_i - Y)^2 \quad (2.3)$$

Another way of describing this property of the mean is to say that the mean *minimises the sum of squares*.

Demonstrating this property is a bit more difficult than it was for the previous one. What follows may be conveniently ignored if one is willing to accept the property without proof.

One relatively simple method for the appropriately initiated is to use the differential calculus, and setting the first derivative of  $\sum_{i=1}^n (X_i - \bar{X})^2$  to zero, derive the definition formula for the mean. Rather than assume the requisite calculus background, however, we provide an algebraic proof that relies on a technique known as *reductio ad absurdum*. *Reductio ad absurdum* means to reduce to an absurdity, and works as a proof by first assuming the direct

<sup>3</sup>It is conventional to denote deviation scores by the *lower-case* version of the name of the original set or variable.

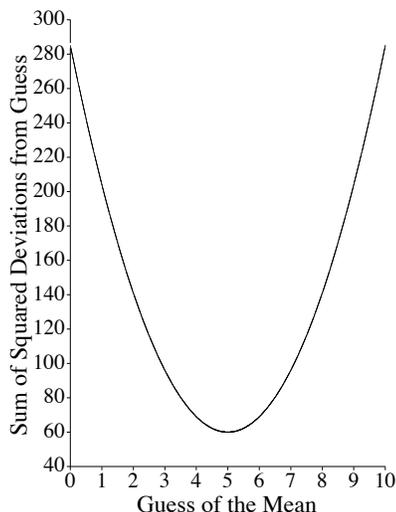


Figure 2.2: The mean minimises the sum of squares. In this example, which plots the sum of squared deviations as a function of different “guesses” of the mean,  $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ ,  $\bar{X} = 5$ , and the minimum sum of squared deviations is 60. Note that this minimum occurs only at the mean.

contradiction of what one wants to prove to be true. If that assumption results in a mathematical absurdity then that assumption *must be false* and, therefore, its contradiction—what we wanted to prove—*must be true*.<sup>4</sup> The argument is presented in Table 2.2

Again, we could use this property of the mean as a method to determine the mean for any distribution, but as it would require searching through an ever-decreasing range of values to find the minimum, it would not be very efficient. Still, it’s comforting to know that if you found yourself one day remembering nothing else about the definition of the mean but this property, you could still home in on its value. On the other hand, if that had happened to you, your concerns undoubtedly would be other than whether or not you could still compute a mean. . .

#### 2.2.4 And the Mean is . . .

All of the foregoing leads to the following definition of the mean: the mean of  $n$  numbers is that number that when added to itself  $n$  times is equal to the sum of the original  $n$  numbers. For example, with  $X = \{1, 2, 3, 4, 5\}$ ,  $\bar{X} = 3$ , and

<sup>4</sup>For *reductio ad absurdum* to work as a mathematical proof requires that the system be *consistent* (i.e., not contain contradictions among its premises or axioms); otherwise, reasoning to an absurdity or contradiction may reflect nothing more than the inherent inconsistency of the system.

We start, then, by assuming that  $\sum_{i=1}^n (X_i - \bar{X})^2$  is *not* the minimum sum, that some *other* value, either larger or smaller than the mean, substituted for the mean, will produce a sum of squared deviations from it *lower* than that obtained with the mean. We represent this value as  $\bar{X} + c$ , where  $c$  is either positive or negative. Then, our contradictory assumption amounts to the statement that:

$$\sum_{i=1}^n (X_i - \bar{X})^2 > \sum_{i=1}^n (X_i - (\bar{X} + c))^2$$

Completing the squares of both sides,

$$\sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n X_i \bar{X} + \sum_{i=1}^n \bar{X}^2 > \sum_{i=1}^n X_i^2 - 2(\bar{X} + c) \sum_{i=1}^n X_i + \sum_{i=1}^n (\bar{X} + c)^2$$

Which yields

$$\begin{aligned} \sum_{i=1}^n X_i^2 - 2 \frac{(\sum_{i=1}^n X_i)^2}{n} + \frac{(\sum_{i=1}^n X_i)^2}{n} \\ > \sum_{i=1}^n X_i^2 - 2(\bar{X} + c) \sum_{i=1}^n X_i + \sum_{i=1}^n (\bar{X}^2 + 2\bar{X}c + c^2) \end{aligned}$$

and

$$\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n} > \sum_{i=1}^n X_i^2 - 2(\bar{X} + c) \sum_{i=1}^n X_i + n\bar{X}^2 + 2n\bar{X}c + nc^2$$

Further reducing the right side of the inequality yields

$$> \sum_{i=1}^n X_i^2 - 2 \frac{(\sum_{i=1}^n X_i)^2}{n} - 2c \sum_{i=1}^n X_i + \frac{(\sum_{i=1}^n X_i)^2}{n} + 2c \sum_{i=1}^n X_i + nc^2$$

resulting in

$$\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n} > \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n} + nc^2$$

Which, because  $nc^2$  has to be positive, is impossible. Therefore, by *reductio ad absurdum*  $\sum_{i=1}^n (X_i - \bar{X})^2$  *must* be the minimum.

---

Table 2.2: Demonstration that the sum of squared deviations from the mean is a minimum.

$\sum_{i=1}^n X_i = 1 + 2 + 3 + 4 + 5 = 3 + 3 + 3 + 3 + 3 = n\bar{X} = 5 * 3 = 15$ . It is in this sense that the *mean* of  $X$  can truly be considered to be *representative* of the numbers in the set: it is the *only* number that can be substituted for each and every number in the set and still result in the same sum. As an aside, it is also for this reason that the mean is frequently substituted for missing data in a set of numbers.

### 2.2.5 The Method of Provisional Means

The “method of provisional means” refers to a recursive method to computing a mean, similar to the recursive method of computing sums described in section A.2.2. It is also referred to as the “running-mean algorithm”, and is commonly used in computer programs (especially in earlier years when computer memory was prohibitively expensive) to compute the mean without having to have all the data simultaneously stored in computer memory.

For example, suppose you wished to take the mean of a mere one-million ( $10^6$ ) values. On many modern computers, each value would require anywhere from 4 to 8 bytes of storage. With a million such values, that would require between  $4 \times 10^6$  and  $8 \times 10^6$  bytes, which, because computer memory is measured in powers of 2, would be between  $(4 \times 10^6)/1024^2 = 3.8$  and  $(8 \times 10^6)/1024^2 = 7.6$  *megabytes* of memory just to store the data. Given that most computer languages pass the data to functions *by value* rather than *by reference*, the memory requirements would double as the data were duplicated to be passed to the function for processing. The solution to this memory problem is simple: don’t store the data in computer memory; rather, store it in a data file or some such, and read in each value, one by one, as needed, discarding it before reading in the next value. But we have to read the data in one pass, otherwise the computation would be markedly slow.

The algorithm is based on the fact that the sum of the scores can be derived from the product of the number of scores,  $n$ , and the mean based on those  $n$  scores,  $\bar{X}_n$ . That is, equation 2.1 may be re-written as:

$$\sum_{i=1}^n X_i = n\bar{X}_n$$

which means that the mean of  $n$  scores may be re-written in terms of the mean for the first  $n - 1$  scores as:

$$\bar{X}_n = \frac{(n-1)\bar{X}_{n-1} + X_n}{n}$$

A little algebra yields:

$$\bar{X}_n = \bar{X}_{n-1} + \frac{X_n - \bar{X}_{n-1}}{n} \quad (2.4)$$

a *recursive* equation for calculating the mean (see Spicer, 1972; Vokey, 1990). That is, the mean for the first  $n$  scores may be calculated from the mean for the first  $n - 1$  scores plus the deviation of the  $n$ th score from the just previous mean divided by  $n$ .

## 2.3 Other Means

The mean provides the *general* location of the distribution, but sometimes fails to provide a value that is truly *representative* of the magnitude of the numbers in the distribution. For example, if the distribution contained many large positive values, and many extremely negative values, with few other values in between, the mean would tend toward zero, balancing the two extremes. Yet, few of the numbers would be anywhere near zero.

### 2.3.1 The Absolute Mean

One straightforward approach to the problem of means failing to be representative of the magnitude of the numbers in the set is to compute the *absolute mean*—the mean of the *absolute value* of the numbers in the distribution or set. Taking the absolute value of a number converts a negative number to its positive equivalent simply by stripping away the negative sign. Positive numbers are unaffected by the operation. The absolute value of a number is denoted by flanking the number with parallel vertical lines. Thus, for example,  $|-3| = |3| = 3$ .

In summation notation, the absolute mean is defined as:

$$\bar{X}_{abs} = \frac{\sum_{i=1}^n |X_i|}{n} \quad (2.5)$$

For the set of numbers,  $X = \{-10, -9, 6, 9, 4\}$ , the mean,  $\bar{X} = 0$ , but the absolute mean,  $\bar{X}_{abs} = 38/5 = 7.6$ , which is noticeably closer in absolute magnitude to the numbers of the set. For example, say you have a choice between two poker games and you want to know how high the stakes are in each. One answer to that question is to compute the average amount of money that changes hands in each game. For game X, imagine that the winnings and losses of the four players were  $X = \{-10, 30, -20, 0\}$ ,  $\bar{X} = 0$ , and for game Y, they were  $Y = \{-1000, 3000, -2000, 0\}$ ,  $\bar{Y} = 0$ . Using the arithmetic mean, then, the two games appear equivalent *on average* (i.e., both are zero-sum games). However, comparing the absolute means,  $\bar{X}_{abs} = (10 + 30 + 20 + 0)/4 = \$15$  vs.  $\bar{Y}_{abs} = (1000 + 3000 + 2000 + 0)/4 = \$1500$ , reveals that the stakes in game Y are much higher. The absolute mean maintains that 100/1 ratio of the size of the number and is thus more informative for you. That's not to say that the arithmetic mean of 0 in each case is wrong, it just doesn't tell you what you want to know—and remember, the whole point of a statistic is to tell you something useful.

### 2.3.2 Root-Mean-Square

Another common approach is to *square* the numbers to remove the sign, compute the mean squared value, and then take the square-root of the mean-square to return to the original units. This value is known as the *root-mean-square* or *r.m.s.* To compute it, you follow the name in reverse (i.e., square each score, take the mean of the squared scores, and then take the square-root of the mean). In summation notation:

$$\bar{X}_{rms} = \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}} \quad (2.6)$$

For example, for the same set of numbers,  $X = \{-10, -9, 6, 9, 4\}$ ,  $\bar{X}_{rms} = \sqrt{= 314/5} = 7.92$ , which is also noticeably closer in absolute magnitude to the numbers of the set. In fact, except where all the numbers are the same,  $\bar{X}_{rms} > \bar{X}_{abs}$ . That is, in general,  $\bar{X}_{rms}$  is always a little larger than  $\bar{X}_{abs}$ .

$\bar{X}_{rms}$  often is used in situations in which the data tend to move back and forth, like a sine wave, AC current, loudspeakers producing sound (hence, its use in stereo system specifications, e.g., 100 watts rms). Put a kid on a swing and measure the endpoints of the swing as it goes back and forth. You'll get something like:  $X = \{10, -10, 11, -11, 9, -9\}$  for going back and forth 3 times; compare that to:  $Y = \{5, -5, 5.5, -5.5, 4.5, -4.5\}$ . Who has swung farther? The arithmetic mean for both is 0, which is not informative, as with the aforementioned poker games. However,

$$\bar{X}_{rms} = \sqrt{(10^2 + (-10)^2 + 11^2 + (-11)^2 + 9^2 + (-9^2))/6} = 10.033$$

and

$$\bar{Y}_{rms} = \sqrt{(5^2 + (-5)^2 + 5.5^2 + (-5.5)^2 + 4.5^2 + (-4.5^2))/6} = 5.016$$

so clearly  $\bar{X}_{rms}$  captures the cyclic peaks (and, hence, the power in the swing or the loudness of the speakers) more meaningfully than does the arithmetic mean.

### 2.3.3 Geometric Mean

The geometric mean takes another approach altogether. It is only defined for positive scores greater than zero, but is still quite useful. It is directly concerned with such things as geometric series (e.g., 2, 4, 8, ...; or 10, 100, 1000, ...) and logarithms, and represents the "central tendency" or "middle" value of such a series of log-transformed data. For a set of  $n$  scores, the geometric mean is that number that when *multiplied* by itself  $n$  times produces the same *product* as the cumulative product of the  $n$  scores in exactly the same way that the arithmetic mean is that number that when *added* to itself  $n$  times produces the same *sum* as the cumulative sum of the  $n$  scores.

For a set,  $X$ , consisting of  $n$  scores, the geometric mean is defined as:

$$\bar{X}_G = \left( \prod_{i=1}^n X_i \right)^{\frac{1}{n}} \quad (2.7)$$

The  $\Pi$  operator denotes the *product* of the operands in the same way that the  $\Sigma$  symbol denotes the sum. Equation 2.7 says that to compute the geometric mean, the scores are cumulatively multiplied together, and then the  $n$ th root of the resulting product is taken. For example, for  $X = \{10, 100, 1000\}$ ,  $\bar{X}_G = 1000000^{\frac{1}{3}} = 100$ , the middle number of the geometric series. Exactly the same result would have been obtained if we had first log-transformed the data (say, to base 10, although the actual base used is not critical), to produce the scores  $\log_{10}(X) = \{1, 2, 3\}$ , and then had taken the anti-log to the same base of the *arithmetic mean* of the transformed scores (i.e.,  $10^2 = 100$ ). If the data are not a geometric series, the geometric mean of the data still returns the (anti-log of the) arithmetic mean of the log-transformed data; in effect, it removes the *base* of the data from the computation of the mean.

Think of it this way. Imagine rewriting each score,  $X_i$ , as a constant,  $b$ , raised to some exponent,  $x_i$ , e.g.,  $\log_b(X) = b^{x_1}, b^{x_2}, \dots, b^{x_n}$  (which is precisely what a  $\log_b$ -transform does). The geometric mean is simply that same constant raised to the arithmetic average of the exponents, i.e.,  $\bar{X}_G = b^{\bar{x}}$ .

The geometric mean is commonly used in such fields as economics, so examples of its use tend to come from that domain. Here is a simple example that emphasises the distinction between the arithmetic and the geometric means. Imagine the following game. There are 10 people in a room, each asked to guess a different number between 1 and 100; the person with the guess most distant from a randomly chosen number for that trial is asked to leave with the cash prize for that trial, and the prize doubles for those remaining. Let's say that the prize starts at \$2. Thus, the first person to drop out wins \$2, and the last person remaining wins  $\$2^{10} = \$1024$ . What is the *average* amount won? Using the arithmetic mean, the value would be  $(2 + 4 + 8 + 16 + 32 + 64 + 128 + 256 + 512 + 1024)/10 = \$204.6$ —among the higher numbers (i.e., it is biased toward the larger values), whereas, the geometric mean comes to \$45.25, midway between the bottom-half and top-half of the scores.

### 2.3.4 Harmonic Mean

Yet another approach is the harmonic mean, so-called because it tends to “harmonize” the scores it “averages”. It is defined as

$$\bar{X}_H = \frac{n}{\sum_{i=1}^n X_i^{-1}} \quad (2.8)$$

In words: it is the reciprocal of the average of the reciprocals. For example, for  $X = \{1, 2, 3, 4, 5\}$ ,  $\bar{X}_H = 5/(1 + 1/2 + 1/3 + 1/4 + 1/5) = 2.19$ . It is most commonly used to find the “average” of a set of differing frequencies or counts in a way that takes into account the *proportional* differences among the frequencies, as, for example, in attempting to “harmonize” differing sample sizes relative to their marginal values. One example of such a set of numbers is given by a set of *rates*, such as kilometres per hour, litres per 100 kilometres, cancer deaths per capita, and so on.

Imagine travelling 200 km to Calgary. For the first quarter of the distance of the trip you travel 100 km/h, and then for the next quarter of the distance because of highway construction, you are slowed to 50 km/h (speed fines double!); in the next quarter, you get to travel at the full 110 km/h allowed for by highway 2. Unfortunately, as you encounter the last quarter of the distance to Calgary, the inevitable traffic jams, stalled cars and trucks, and the general freeway insanity that is Calgary, slows you to 10 km/h (and you swear you will never go to Calgary again). Given all that, when you finally arrive, your friendly, Calgary host asks you innocently, what was your average speed? You might think that the appropriate answer is just the arithmetic mean of  $X = \{100, 50, 110, 10\}$ , or  $\bar{X} = 67.5$ —that is, you averaged 67.5 km/h. But note, the trip took 30 minutes to cover the first 50 km, 60 minutes to cover the second 50 km, 27.27 minutes to cover the third 50 km, and 300 minutes to cover the last 50 km, for a total of 417.27 minutes, or 6.95 hours! If you really did average 67.5 km/h, then in 6.95 hours, you would have travelled  $6.95 \times 67.5 = 469.13$  km! But the whole trip was only 200 km! The harmonic mean, in contrast,  $\bar{X}_H = 4/(1/100 + 1/50 + 1/110 + 1/10) = 28.758$ , when multiplied by the number of hours, yields the correct distance of 200 km.

## 2.4 The Median

The *median* is an alternative to the mean to locate the set or distribution of scores along the number line. Because the mean emphasizes the interval properties (i.e., the quantitative differences) of the numbers, and *balances* them, its value can be influenced, sometimes quite markedly, by extreme values or *outliers*. For example, in the set  $X = \{1, 2, 3, 4, 100\}$ ,  $\bar{X} = 110/5 = 22$ , a value unlike the majority of the numbers in the set. The median, on the other hand, is concerned only with the *ordinal* properties of the numbers in the set, and not any other differences among them. Thus, for example, the value of 100 in the set  $X = \{1, 2, 3, 4, 100\}$  has the same ordinal value, 5, as do the numbers 5, 17, 93, 1103, etc. if substituted for it. Consequently, extreme values have less of an effect on the median than they do on the mean.

### 2.4.1 Definition of the Median

The median is defined as a value mid-point between the 50% of the scores below it and the 50% of the scores above it when the scores are arranged in rank-order. As such, it is often referred to as the *50th percentile*. So, in the set  $X = \{1, 2, 3, 4, 100\}$ , for which  $n$  is odd, the median, commonly abbreviated as *mdn*, would be 3. In a set for which  $n$  is even, such as  $X = \{1, 2, 3, 4, 5, 100\}$ , the median could be any value between the two middle values, 3 and 4, such as 3.1 or 3.936 (as any such value would meet the definition), although it is conventional simply to report the average of the two flanking values, in this case 3.5. Note that in this case the median could also be defined as the number at ordinal position  $.5(n + 1)$ .

Implicit in this definition of the median is a more formal definition. Because 50% of the scores are above the median and 50% are below, the following expression for even<sup>5</sup>  $n$  (and all scores different) is true if  $mdn$  really is the median:

$$\sum_{i=1}^n \frac{X_i - mdn}{|X_i - mdn|} = 0 \quad (2.9)$$

All this expression says that if you label each score with +1 if it is above the median, and with -1 if it is below the median, the number of +1 scores will precisely balance the number of -1 scores, for a grand sum of zero.

### 2.4.2 A complication

Sets containing many duplicates of the middle value can be problematic in meeting this definition of the median, although this difficulty is often ignored, and with good reason. First, the approach shown here complicates what is otherwise a straightforward and simple procedure for determining a median. Second, for many distributions, the simply-determined median value is also the modal (i.e., most frequent) value, and is often worth retaining for that reason. Third, for many distributions, especially those collected along continuous scales, the possibility of identical values rarely arises. Fourth, the approach shown is most useful for situations where different values have been “binned” or collected into intervals. For these reasons, this subsection can probably be ignored with little loss.

Take the set  $X = \{1, 2, 3, 3, 3, 3, 3, 3, 4, 5, 5, 5\}$ , for example. It is clear that the median should be somewhere among all the duplicated threes, but simply settling for 3 as the median does not seem to be in the spirit of the definition in that only  $4/12 = 33.3\%$ , the 4 and fives, rather than 50% of the numbers in the set are greater than 3, and only  $2/12 = 16.7\%$ , the 1 and 2, are less than 3. In fact, as there are 12 numbers (i.e.,  $n = 12$ ) in the set, then the middle or 50th percentile number should be the number with a rank-order of  $.50(n + 1) = 6.5$ —between the fourth and fifth 3 in the example set.

For this location of the median to be meaningful requires that the duplicate values themselves be rank-ordered in some way; the 3 closest to the 4 in the series, for example, be *in some sense* a greater value of 3 than the one closest to the 2. How could this be? Assuming that the values in the set were not true integers (as would be the case if they were counts, for example), but instead were obtained as the result of some kind of measurement on a continuous scale, then each of the threes in the set could be seen as *rounded approximations* to the true measured values. Following the conventional procedures for rounding numbers, this assumption implies that the “different” threes in the example set were obtained from measurements ranging in value from a low of 2.5 to a high of 3.4999... , and could be rank-ordered by these assumed *true* values. Thus, the median for this set would be located in the interval 2.5 to 3.4999... . But where,

<sup>5</sup>The expression is true for odd  $n$  as long as the summation is over all values of  $X$  *except* the median to avoid the division by zero problem that would otherwise result.

exactly? We know that it should be located at a value with a rank-order of 6.5. The numbers 1 and 2 already occupy the first and second ordinal positions, the 4 and the fives occupy the ninth through twelfth positions, and the threes occupy the positions 3 through 8. So, we want the median to be located  $6 - 2 = 4$  ordinal positions into the interval occupied by the threes. As the interval contains 6 threes (or is 6 threes wide) in this set, then we want the median to be located  $4/6 = .667$  (i.e., more than one-half) of the way into the interval, for a median of  $2.5 + .667 = 3.167$ .

In general, this definition for the median, which is most useful for *grouped-frequency distributions* (i.e., where the data have been “binned” into intervals of width  $w$ ), may be written as:

$$mdn = l + \left( \frac{.5n - s}{f} \right) w \quad (2.10)$$

In which

$l$  = the lower real limit of the interval (2.5 in our example case)

$n$  = total number of cases (12 in the example)

$s$  = total number of cases below the critical interval (2 in the example)

$f$  = frequency of cases in the critical interval (6 in the example)

$w$  = interval size (1 in the example)

Other percentile statistics may be similarly determined. For example, the 90th percentile is simply that number such that 90% of the scores are below it, and 10% are above it.

### 2.4.3 Properties of the Median

The median has a property similar to that of the mean. Whereas the mean minimises the sum of squared deviations from it, the median (narrowly defined) minimises the sum of *absolute deviations* from it. That is, the median could be

defined as that value for which  $\sum_{i=1}^n |X_i - mdn|$  is a minimum. This definition of

the median is more restrictive than the one offered in section 2.4.1, and is rarely invoked. It does, however, provide for a relationship between the values of the mean and median of a distribution of scores that proves useful for the detection of *skewness*, discussed in Chapter 5, section 5.1.1.

## 2.5 Questions

1. The sum of a set of scores equals 90, with a mean of 10. How many scores are in the set?

2. For a set of 19 scores with a mean of 11, the following sum is computed:  
$$\sum_{i=1}^{19} (X_i - q) = 0.$$
 What is the value of the constant  $q$ ?
3. For your part-time work over 12 months, you received 12 cheques. Although the amount of each cheque was different from all of the others, the average was \$500. How much did you make in total?
4. On a recent quiz, the mean score for the 10 males in the class was 70, and for the 20 females was 80. What is the mean for the class as a whole?
5. If it were not for the low grade of Allen Scott—the class ne'er-do-well (and the only male in the class with two first names)—the mean grade for the 5 males in the class would have equalled that of the 10 females, who had a mean grade of 90. The mean for the class as a whole was 85. What was Allen Scott's grade?
6. In a set of 12 scores with a mean of 60, the sum of deviations from the mean of the 10 scores *below* the mean is found to equal -120. One of the remaining two scores is equal to 85. What is the value of the other one?
7. For the set of scores,  $X = \{2, 5, 1, -4, 12, -7, -2\}$ , compute the mean, the absolute mean, and the root-mean-square mean.
8. In a set of scores, the mean of the 18 males equalled the overall mean of 27. What was the sum of the 34 female scores?
9. Each of the 10 individual investments that had a mean of \$300 in the Vokey benevolent fund increased by 50%. What is the mean of the current investment?
10. A set of 60 scores is known to have a mean of 56. Unfortunately, 9 of the scores are now missing, and the remaining scores have a mean of 50. What is the mean of the missing 9 scores?
11. For reasons not clear even to himself, Allen Scott finds that if he adds 2.756 to every one of his 10 course grades after multiplying each of them by 47, he almost precisely matches the distances in astronomical units (*Earth = 1au*) of the 10 planets of the solar system.<sup>6</sup> If the sum of the course grades was 22.8, what is the mean of his estimates of the planetary distances?
12. The mean of 10 scores is 6. If these scores are in fact the logarithms to the base 2 of the original 10 scores, what is the geometric mean of the original scores?

---

<sup>6</sup>In his dreams; there is nothing about this problem that relates directly or indirectly to the planetary distances of the solar system in *au* or otherwise.

13. The mean of 23 numbers is found to be 38. If a twenty-fourth number of 16 is to be included in the mean, what would that new mean be?
14. A distribution of 10 scores has one score that is very much greater than all the others. Both the mean and median are calculated. Which one is likely to be smaller?
15. The students in a statistics (sic) class report that they spent a total of \$900 on textbooks and calculators for the course, a mean of \$45 per person. One-quarter of the class is male, and males spent a mean of \$60 for their books and calculators.
  - (a) How many students are in the class?
  - (b) What was the mean amount spent by females?
16. You run an experiment with a control group and an experimental group in which the mean for the 3 scores in the control group was 20; the mean for the experiment as a whole, which included the 5 scores from the experimental group, was 22.5. Although the results suggest a difference between the two groups, you are concerned because *without* the lowest score in the control group, the control mean would have equaled the mean for the experimental group. What was this lowest score?



## Chapter 3

# Measures of Variability

After the *location* of the data, the next property of interest is how *variable* or *dispersed*—how spread out—the data are. Two sets or distributions of numbers could be otherwise identical (e.g., same mean, same median, same overall shape, etc.), but still differ in how variable or *different from one another* the numbers in each set are. There have been many different measures of variability proposed and used over the years, reflecting different but related aspects of a distribution of numbers. We will highlight only a few. These few are selected for discussion because either they are commonly used in experimental psychology, or because they have properties that are exploited in subsequent statistics, or typically both.

### 3.1 The Range

The *range* is probably the simplest measure of dispersion, and for a *PDQ*<sup>1</sup> measure of spread, it is often quite useful. The range is defined simply as the highest score minus the lowest score; the intermediate scores are otherwise ignored. For example, for the set of scores  $X = \{1, 8, 9, 17, 25, 47\}$ , the range =  $47 - 1 = 46$ .

Because the range uses only the highest and lowest values and ignores the remaining numbers (except for the fact that they are intermediate), it throws away much of the information about the distribution. Two distributions could have the same range, but otherwise be quite different in how variable the remaining scores were. Consequently, for many purposes, it is thought preferable to use a measure that directly reflects the differences between *every* number and every other number.

In any set of numbers, pairing every number with every other number, including itself, results in  $n * n = n^2$  pairs of numbers and, hence,  $n^2$  differences among them. Perhaps combining these differences in some way will produce a

---

<sup>1</sup>“Pretty Damn Quick”—an important measure of utility.

summary measure of variability or inter-item differences. Unfortunately, simply summing or averaging these differences does not result in anything useful because the sum is always zero. This result is easily seen. First, obviously the  $n$  differences of pairs of each number with itself sum to zero as each difference is zero. For the remaining pairs, every pair difference of  $a - b$  is directly counterbalanced by the corresponding reversal of the pair difference, or  $b - a$ , so again the sum is zero. More formally, in summation notation:

$$\sum_{i=1}^n \sum_{j=1}^n (X_i - X_j) = 0$$

But because  $X_i$  is a constant for the sum over  $j$ , and likewise  $X_j$  is a constant for the sum over  $i$ , we get

$$n \sum_{i=1}^n X_i - n \sum_{j=1}^n X_j = 0$$

### 3.2 The $D^2$ and $D$ statistics

To counteract the problem of the positive  $a - b$  differences cancelling the corresponding negative  $b - a$  differences, we could *square* each difference and then sum and take the mean, as we did for the  $\overline{X}_{rms}$  statistic. We'll define  $D^2$ —the average squared difference over all possible  $n^2$  pairs, then, as:

$$D^2 = \frac{\sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2}{n^2} \quad (3.1)$$

And we could again take the positive square-root to get back to the original units, called  $D$ :

$$D = \sqrt{\frac{\sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2}{n^2}} \quad (3.2)$$

Unfortunately, although it does use all the scores, and does return a value  $\geq 0$ , it is a tedious procedure, even for small  $n$ . For example, for 8 numbers, there are 64 pairs, for 10, 100 pairs, for 20, 400 pairs, and so on. And it gets worse and worse, faster and faster,

### 3.3 Variance ( $S^2$ ) and Standard Deviation ( $S$ )

Fortunately, there is another way that is related to the calculation of  $D^2$  and  $D$ , but this time involves the mean. Whereas  $D^2$  computed the squared distance between each item and every other item directly,  $S^2$  does so indirectly. The

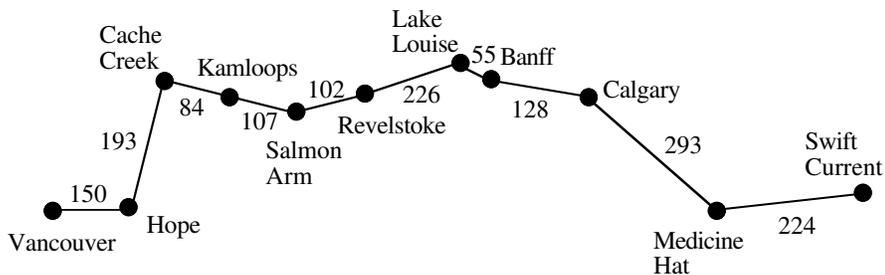


Figure 3.1: A map of distances between towns and cities along the Trans-Canada highway.

essential idea is as follows. Shown in Figure 3.1 is a simplified map of towns and cities along the Trans-Canada highway from Swift Current in the East to Vancouver in the West. To specify the distance between every town and city and every other town and city we could, as we did with  $D^2$ , construct a square table or *matrix* of  $n^2$  cells (where  $n$  is the number of towns and cities) that lists every town or city as both a column and a row, with the distance between each row town or city and each column town or city as a value in the cell formed by their intersection in the table. The cells along the diagonal from the top left to the bottom right would be zeros—representing the distance between each town or city and itself—and the off-diagonal cells would contain the distances between the row town or city and the column town or city. Such a matrix of the distances on the map is shown in Table 3.1.

But we also could represent the same information more compactly by choosing *one town or city* or even some *arbitrary point* as the default or standard, and simply record the distance of every town or city from it. Even though we wouldn't have directly represented them, it is the case that the distances between each town or city and every other town or city are indirectly represented in this format. As Calgary is the centre of the known universe—at least to Calgarians—let's choose it as our focal Trans-Canada city, as shown in Table 3.2.

For example, as Kamloops is 618 km from Calgary, and Medicine Hat is  $-293$  km (the sign of the values representing which side of the default city the other cities or towns are), then the distance between Kamloops and Medicine Hat is  $618 - (-293) = 911$  km. Thus, by recording only the distance between every town and city and the default of Calgary, we manage to retain indirectly the distances between every town and city and every other town and city, reducing the number of distances to compute from  $n^2$  to  $n$ .

$S^2$  takes this latter approach to summarising the difference between every number and every other number in a set. It uses the mean as its default point, and is computed as the average squared difference from the mean:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \quad (3.3)$$

	Van.	Hope	Cache Cr.	Kam.	Sal. Arm	Rev.	Lk. Lou.	Banff	Cal.	Med. Hat	Swift Cur.
Vancover	0	150	343	427	534	636	862	917	1045	1338	1562
Hope	150	0	193	277	384	486	712	767	895	1188	1412
Cache Creek	343	193	0	84	191	293	519	574	702	995	1219
Kanloops	427	277	84	0	107	209	435	490	618	911	1135
Salmon Arm	534	384	191	107	0	102	328	383	511	804	1028
Revelstoke	636	486	293	209	102	0	226	281	409	702	926
Lake Louise	862	712	519	435	328	226	0	55	183	476	700
Banff	917	767	574	490	383	281	55	0	128	421	645
Calgary	1045	895	702	618	511	409	183	0	128	293	517
Medicine Hat	1338	1188	995	911	804	702	476	421	293	0	224
Swift Current	1562	1412	1219	1135	1028	926	700	645	517	224	0

Table 3.1: Distances (in km) from each city or town shown in Figure 3.1 to every other city or town shown on that map. The cities are listed in order from West to East. Note that each place is 0 km from itself (the diagonal) and that each distance is displayed twice: once in the upper-right triangle and once in the lower-left triangle.

	distance from Calgary
Vancouver	1045
Hope	895
Cache Creek	702
Kamloops	618
Salmon Arm	511
Revelstoke	409
Lake Louise	183
Banff	128
Calgary	0
Medicine Hat	-293
Swift Current	-517

Table 3.2: Distances (in km) from each city or town shown in Figure 3.1 to Calgary. Note that using Calgary as a reference point and defining distances eastward as negative and westward as positive (both typical Calgarian biases), allows for a substantially reduced table. Any distance shown in Table 3.1 can be calculated from the information in this table requiring a maximum of 1 subtraction (e.g, the distance from Salmon Arm to Banff equals  $511 - 128 = 383$  km as shown in Table 3.1).

$S^2$  is a *mean-square*; it is referred to as *the variance* of a distribution because it captures the *variability* or dispersion of the numbers. The square-root of the variance,  $S$ , is called *the standard deviation*, and is just the variability expressed in the original units of the distribution:

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}} \quad (3.4)$$

The choice of the mean as the default point from which to compute deviations is not arbitrary. As the mean minimises the sum of squared deviations from it (see section 2.2.3),  $S^2$  for any distribution is *as small as it can be*. Hence, if  $S_X^2$  for one distribution  $X$  is greater than  $S_Y^2$  for another distribution  $Y$  then because both are as small as they can be, the difference between them must reflect the fact that distribution  $X$  is more variable than distribution  $Y$ .

In the definitional form of equation 3.3,  $S^2$  is still tedious to calculate. We can algebraically manipulate it to produce a more formidable looking, but easier to use *computational formula* for hand calculation (see Table 3.3 for the derivation):

$$S^2 = \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - \frac{\left( \sum_{i=1}^n X_i \right)^2}{n} \right) \quad (3.5)$$

The emphasis on *hand calculation* is important. “Computational formulae” are rarely of much use as computer algorithms or spreadsheet formulae. Often they are slower than other alternatives, and, as in the case of the computational formula for variance, have very poor numeric precision. Computers store numbers to a limited numeric precision, which varies with computer-type and operating system, although many use a standard of 6 to 8 digits of precision. Thus, very large and very small numbers are poorly represented because of the rounding necessary to pack the numbers into 6 to 8 digits.<sup>2</sup> Thus, computations that accumulate very large values, such as the large sums of squares and squared sums in the computational and mnemonic formulae for variance, can result in significant rounding error, especially as  $n$  increases. Hence, although such formulae are useful for small sets of numbers and the hand calculations on them, they *should not* be used for any more elaborate computations. Unfortunately, some commercially-available statistics programs and functions in spreadsheets are still based on these formulae, rendering large-scale analyses with these programs unreliable.

If we continue the algebraic manipulation by multiplying the fraction  $1/n$  through,

$$S^2 = \frac{\sum_{i=1}^n X_i^2}{n} - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n^2}$$

we obtain the surprisingly simple *mnemonic formula* for variance:

$$S^2 = \overline{X^2} - \bar{X}^2 \tag{3.6}$$

the *mean-square minus the squared-mean*. Also surprising is the fact that, despite computing the deviations from the mean of the distribution rather than from every other score,  $S^2$  is really just a (much) more efficient way of computing  $D^2$ . In fact,  $D^2 = 2S^2$  (see Table 3.4 for the derivation). So,  $S^2$  or the variance does reflect the average squared difference of the scores from one another in addition to reflecting the average squared difference from the mean. With the exception of a constant factor of 2, the concepts are interchangeable.

Many hand calculators, computer statistical applications, and computer spreadsheet applications provide functions for variance and standard deviation. Unfortunately, these functions sometimes compute what is known as a population variance (or standard deviation) *estimate* (see Chapter 14, section 14.1 for further information on this concept). As not all calculators, statistical applications, and spreadsheets necessarily default to one or the other calculation—and few document exactly how the function is calculated, you may have to test the function to see whether it is dividing by  $n$ —the correct value—or by  $n - 1$  to produce a population estimate. To do so, test the function with the values 1,

---

<sup>2</sup>A related approach is to use a fixed number of *bits* of internal representation and encode each number in a *floating-point* form (i.e., the base-2 equivalent of exponential notation). The effect of limited precision, however, is similar.

$$\begin{aligned}
S^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \\
&= \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \right) \\
&= \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - 2 \frac{\sum_{i=1}^n X_i}{n} \sum_{i=1}^n X_i + n \left( \frac{\sum_{i=1}^n X_i}{n} \right)^2 \right) \\
&= \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - 2 \frac{\left( \sum_{i=1}^n X_i \right)^2}{n} + \frac{\left( \sum_{i=1}^n X_i \right)^2}{n} \right) \\
&= \frac{1}{n} \left( \sum_{i=1}^n X_i^2 - \frac{\left( \sum_{i=1}^n X_i \right)^2}{n} \right)
\end{aligned}$$

Table 3.3: Derivation of the computational formula for variance.

$$\begin{aligned}
D^2 &= \frac{\sum_{i,j=1}^n (X_i - X_j)^2}{n^2} \\
&= \frac{1}{n^2} \sum_{i,j=1}^n (X_i^2 - 2X_iX_j + X_j^2) \\
&= \frac{1}{n^2} \left( \sum_{i,j=1}^n X_i^2 - 2 \sum_{i,j=1}^n X_iX_j + \sum_{i,j=1}^n X_j^2 \right) \\
&= \frac{1}{n^2} \left( n \sum_{i=1}^n X_i^2 - 2 \left( \sum_{i=1}^n X_i \right)^2 + n \sum_{j=1}^n X_j^2 \right) \\
&= \frac{\sum_{i=1}^n X_i^2}{n} - 2 \frac{\left( \sum_{i=1}^n X_i \right)^2}{n^2} + \frac{\sum_{i=1}^n X_i^2}{n} \\
&= 2 \left( \frac{\sum_{i=1}^n X_i^2}{n} - \frac{\left( \sum_{i=1}^n X_i \right)^2}{n^2} \right) \\
&= 2S^2
\end{aligned}$$

Table 3.4: Demonstration of the equivalence of  $D^2 = 2S^2$ .

2, and 3. If the value it returns is  $\sqrt{0.67} = 0.82$ , then it is dividing by  $n$ ; if it returns 1.0, then it is dividing by  $n - 1$ , and must be corrected. Accordingly, one can correct the result provided by the built-in function. If the function computes the standard deviation by dividing the sum of squared deviations by  $n - 1$  rather than  $n$ , the correction, then, consists of multiplying the value returned by the function by  $\sqrt{(n - 1)/n}$  to get the correct standard deviation of the scores.

### 3.4 Questions

1. For a set 20 scores, the mean is found to be 10, and the sum of the squared scores is found to be 2500. What is the standard deviation of the 20 scores?
2. Compute the standard deviation of  $X = \{3, 5, 8, 2, 9\}$  *without* first computing the mean.
3. For a set of scores, if  $S_X^2 = 20$ ,  $\bar{X} = 10$ , and the sum of the squared score equals 600, how many scores are in the set?



## Chapter 4

# Transformed Scores

Often it is desirable to transform the data in some way to achieve a distribution with “nicer” properties so as to facilitate comparison with some theoretical distribution or some other set of data. Such transformations can be especially useful for data-sets for which the actual values of the numbers are of little or no intrinsic interest. Common examples of such transformations or *re-scaling* are found in IQ, GRE, and SAT scores, grade-point-averages, course-grades computed by “grading on a curve”, temperatures (e.g., Fahrenheit, Celsius, or Absolute scales), currencies, and so on.

### 4.1 The Linear Transform

Although transformations of any mathematical type are possible, the most common are *linear* transformations. Linear transformations are those that accomplish the transformation through the use of the addition (or subtraction) and/or the multiplication (or division) of constants. Generically, a linear transformation of  $X$  is given as:

$$Z_i = a + bX_i$$

where  $a$  and  $b$  are constants. It is called a linear transform because if the transformed value  $Z$  is plotted as a function of the original value  $X$ , a straight line is produced.<sup>1</sup>

#### 4.1.1 Rules for changing $\bar{X}$ , $S_X^2$ and $S_X$

Linear transformations of data are often used because they have the effect of adjusting the mean and variance of the distribution.

---

<sup>1</sup>Technically, the transform we are referring to here is known in mathematics as an *affine* transform. In mathematics, a linear transform is any transform  $T$  that for every  $X_1$  and  $X_2$  and any number  $c$ ,  $T(X_1) + T(X_2) = T(X_1 + X_2)$  and  $T(cX_i) = cT(X_i)$ . So, an affine transform is a linear transform plus a constant.

**Adding a constant to every score**

Adding a constant (negative or positive) to every score in a distribution has the effect of shifting the distribution as a whole up or down the number line by the amount of the constant. It does not effect the *shape* of the distribution nor the difference between any score and any other score. Hence, it should affect the mean by shifting it the same amount as the constant, but have no effect on the variance or standard deviation. Formally, if

$$Z_i = X_i + c$$

then, for the mean:

$$\begin{aligned}\bar{Z} &= \frac{\sum_{i=1}^n (X_i + c)}{n} \\ &= \frac{\sum_{i=1}^n X_i + nc}{n} \\ &= \frac{\sum_{i=1}^n X_i}{n} + \frac{nc}{n} \\ &= \bar{X} + c\end{aligned}$$

For the variance,

$$\begin{aligned}S_Z^2 &= \frac{\sum_{i=1}^n (Z_i - \bar{Z})^2}{n} \\ &= \frac{\sum_{i=1}^n ((X_i + c) - (\bar{X} + c))^2}{n} \\ &= \frac{\sum_{i=1}^n (X_i + c - \bar{X} - c)^2}{n} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \\ &= S_X^2\end{aligned}$$

And, hence,  $S_Z = S_X$ .

**Multiplying each score by a constant**

The effects of multiplying (or dividing) each score by a constant are perhaps less obvious. Doubling each score, for example, would have the effect (for all positive scores) of moving the scores up the number line (as each score is now twice as large), but it also would have the effect of doubling the differences between the scores:  $X = \{1, 2, 3\}$  would become  $Z = \{2, 4, 6\}$ . So, both the mean and variance (and standard deviation) should be affected. Formally, if

$$Z_i = cX_i$$

then, for the mean:

$$\begin{aligned}\bar{Z} &= \frac{\sum_{i=1}^n cX_i}{n} \\ &= c \frac{\sum_{i=1}^n X_i}{n} \\ &= c\bar{X}\end{aligned}$$

For the variance,

$$\begin{aligned}S_Z^2 &= \frac{\sum_{i=1}^n (Z_i - \bar{Z})^2}{n} \\ &= \frac{\sum_{i=1}^n (cX_i - c\bar{X})^2}{n} \\ &= \frac{\sum_{i=1}^n c^2 (X_i - \bar{X})^2}{n} \\ &= c^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \\ &= c^2 S_X^2\end{aligned}$$

And, hence,  $S_Z = cS_X$ .

In summary, the mean behaves in exactly the same way as any score in the distribution. Adding or multiplying every score in a distribution by a constant means that the mean for the transformed distribution is obtained by adding or multiplying the mean for the original distribution by the same constant. It is in this sense that the mean can be considered to be a *representative* score of the distribution. The variance and standard deviation are unaffected by additive

constants, but are changed by multiplicative constants: the transformed variance is simply the original variance multiplied by the square of the constant, and the transformed standard deviation is the equal to the original standard deviation multiplied by the constant.

## 4.2 The Standard Score Transform

The standard score transform is, as its name suggests, a transformation that re-expresses any set of scores in a standard format. It is a linear transform that expresses the scores in terms of units of the standard deviation of the distribution. It is defined as:

$$Z_i = \frac{X_i - \bar{X}}{S} \quad (4.1)$$

and is simply the signed number of standard deviations the score is above (positive) or below (negative) the mean of the distribution. A score of 70 from a distribution with a mean of 80 and a standard deviation of 5, for example, would have a standard-score or *Z-score* (as they are often called) of  $-2.0$  because it is 2 standard deviations below the mean of its distribution. A score of 82.5 would have a *Z-score* of 0.5, and so on.

### 4.2.1 Properties of Standard Scores

Transforming every score in a distribution to *Z-scores* produces a transformed distribution with unique, indeed desirable, properties. As noted, a linear transform does not affect the shape of the distribution only, potentially, its location and variability. By subtracting the mean from each score and dividing by the standard deviation, a distribution is produced that has a mean equal to zero and a standard deviation equal to 1. Because  $1^2 = 1$ , this latter property also means that the standard deviation of standard scores equals their variance. More formally, for the mean of a standardised distribution,

$$\begin{aligned} \bar{Z} &= \frac{1}{n} \sum_{i=1}^n \frac{X_i - \bar{X}}{S} \\ &= \frac{1}{nS} \sum_{i=1}^n (X_i - \bar{X}) \\ \text{But as } \sum_{i=1}^n (X_i - \bar{X}) &= 0 \\ &= \frac{1}{nS}(0) \\ &= 0 \end{aligned}$$

For the variance of a standardised distribution,

$$\begin{aligned}
 S_Z^2 &= \frac{\sum_{i=1}^n (Z_i - \bar{Z})^2}{n} \\
 &= \frac{\sum_{i=1}^n Z_i^2}{n} \\
 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \\
 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
 &= \frac{n}{n} \\
 &= 1
 \end{aligned}$$

Incidentally, this derivation also demonstrates that

$$\sum_{i=1}^n Z_i^2 = n$$

### 4.2.2 Uses of Standard Scores

Transforming every score of a distribution to standard scores allows for the easy comparison of a given score from one distribution to that of another in terms of their *relative* difference from their respective means. A score of 16 from a distribution with a mean of 12 and a standard deviation of 1—corresponding to a standard score of 4—can be seen to be more “deviant” in this relative sense than a score of 80 from a distribution with a mean of 50 and a standard deviation of 10—corresponding to a standard score of only 3.

The standard score transform also provides for the easy conversion of scores from one distribution to those of another. For example, scores on such examinations as the Graduate Record Exam (GRE) are transformed to a scale with

a mean of 500 and a standard deviation of 100 via the standard score transform. That is, whatever the mean and standard deviation of the actual scores on the examination, the scores are converted to GRE scores via the following transformation, after first having been converted to standard ( $Z$ -) scores:

$$GRE_i = 100Z_i + 500$$

So, a GRE score less than 500 is automatically recognised as a score less than the mean, and a score of 653 as 1.53 standard deviations above the mean. Scores obtained on “standardised” IQ tests are similarly transformed, although in this case so that the mean is 100 with a standard deviation of 15.

From a statistical standpoint, however, standard scores are quite useful in their own right. As they have a mean of 0 produced by subtracting the raw-score mean from every raw score, they are “centred”, and subsequent calculations with them can occur without regard to their mean. Standardised scores in some sense express the “true essence” of each score beyond that of the common properties (mean and standard deviation) of their distribution. Other advantages are discussed in subsequent sections.

### 4.3 Questions

1. For a set of standardised scores, the sum of squared scores is found to equal 10. How many scores are there?
2. For a set of standardised scores, each score is multiplied by 9 and then has five added to it. What are the mean and standard deviation of the new scores?
3. Express a score of 22 from a distribution of scores with a mean of 28 and a standard deviation of 3 as a linearly transformed score in a distribution with a mean of -10 and a standard deviation of 0.25.
4. On a recent test, Sam—the class ne'er-do-well—achieved a score of 50, which was 3 standard deviations below the class mean. Sandy—the class wizard—achieved a score of 95, which was 2 standard deviations above the class mean. What was the class mean?
5. T-scores have a mean of 50 and a standard deviation of 10. Express a GRE score of 450 as a T-score.
6. The WAIS (Wechsler Adult Intelligence Scale—or so it is alleged) is scaled to have a mean of 100 and a standard deviation of 15. Your dog's score of 74 on the meat-lovers' sub-scale was rescaled to give a WAIS meat-lovers' IQ of 70. If the standard deviation on the meat-lovers' sub-scale is 7, what is the mean?
7. Following a standard score transform, a score of 26 from distribution  $X$  is found to be equivalent to a score of 75 in distribution  $Y$  for which the mean

- is 60 and the standard deviation is 20. If the mean in the  $X$ -distribution is 20, what is the standard deviation of distribution  $X$ ?
8. Larry was given the Transformed Weinberger Intelligence Test, which has a mean of 18 and a standard deviation of 2. He scored 22. His brother Darryl refused, but his other brother Darryl scored 57 on the General Intelligence Mental Profile that has a mean of 50 and a standard deviation of 3. Based on the results of the tests, decide which brother is the more intelligent, explaining why you came to that decision.
  9. IQ scores typically are standardized to have a mean of 100 and a standard deviation of 15. GRE scores are standardized to have a mean of 500 and a standard deviation of 100. T-scores are standardized to have a mean of 50 and a standard deviation of 10. Accordingly, a T-score of 35 would be equivalent to what as a GRE score, and what as an IQ score?
  10. Fred achieved a standard score of -2.5 on a test with a mean of 20 and a standard deviation of 2. Later, feeling sorry for everyone, the instructor added 10 to each score and then doubles them. What were Fred's new score and new standard score?



## Chapter 5

# Other Descriptive Statistics

Reducing a distribution of numbers to its mean and variance may not capture all that is interesting or useful to know about: two distributions may have identical means and variances, for example, but still be quite different in shape. As with the mean and variance, the other aspects of the distribution that underlie these differences in shape can be captured in various, carefully chosen statistics.

### 5.1 Other Statistics

The mean is a statistic such that the sum of deviations from it is zero. That is, it is based on the sum of the deviation scores raised to the first power, or what is known as the *first moment* of the distribution, and captures the location of the distribution. The variance is derived from the sum of the deviation scores raised to the second power, or the *second moment* of the distribution, and captures the variability of the scores. Other statistics can be based on the higher-order moments of the distribution.

#### 5.1.1 Skewness

*Skewness*, for example, is based on the *third moment* of the distribution, or the sum of cubic deviations from the mean, and is often defined as:

$$G_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{nS_X^3}$$

It measures deviations from perfect symmetry. At least one reason it does so is because it uses an odd-numbered exponent. Odd-numbered exponents retain the *sign* of the number or, in the case of deviations from the mean, the sign of

the deviation. Positive skewness indicates a distribution with a heavier positive (right-hand) tail than a symmetrical distribution would have, negative skewness indicates a distribution with a heavier negative tail, and zero skewness indicates perfect symmetry.

### 5.1.2 Kurtosis

*Kurtosis* is derived from the fourth moment (i.e., the sum of quartic deviations), and is often defined as:

$$G_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{nS_X^4} - 3$$

It captures the “heaviness” or weight of the tails relative to the centre of the distribution. One reason it does so is that it uses a larger exponent than does the computation of variance. The larger the exponent, the greater the role large deviations, and thereby tail values, play in the sum. Thus, the resulting statistic deemphasizes small deviations (i.e., values around the mean) in favor of large deviations, or the values in the tails. Hence, although both variance and kurtosis concern deviations from the mean, kurtosis is concerned principally with the tail values. Positive values of kurtosis indicate relatively heavier tails, negative values indicate relatively sparse tails, and a value of zero kurtosis indicates that the tails are about the weight they should be.

Relative to *what?*, you might ask. The answer is relative to the “normal” or Gaussian distribution—the “bell-curve” of many discussions, for example, about the distribution of IQ in human populations. This standard of the normal distribution also explains the “−3” in the equation for the kurtosis statistic. By subtracting 3, the sum equals zero when the distribution in question is “normal”. It also explains why the different sums are “normalised” by the standard deviation ( $S_X$ ) raised to the power of the moment. Such normalisation renders the Gaussian distribution the standard or “normal” distribution: deviations from its shape appear as non-zero values in these equations, with the extent of the deviation in shape from the normal distribution reflected in increasingly deviant values. It is not just that the normal distribution is taken as the norm by fiat or general consensus. Rather, unlike many data distributions, a normal distribution is completely determined by the first two moments. If you know the mean and standard deviation of a normal distribution, you know *everything* there is to know about it. Similarly, if some data-set is well-approximated with a normal distribution, then knowing the mean and standard deviation of that data-set provides virtually all you need to know about that data-set.

Three kurtotic shapes of distributions are generally recognized with labels. “Normal” distributions, having tails of about the “right” weight—neither too heavy or too light, are referred to as “mesokurtic” from “meso” for “middle”. Distributions that are too light in the tails (i.e., very peaked distributions) are called “leptokurtic”, and distributions with too heavy tails (i.e., flattened

distributions) are called “platykurtic”, after the Australian platypus—the first specimens of which had just been displayed in Victorian England at the time that Galton was inventing the names for different degrees of kurtosis.

Clearly, similarly-constructed statistics based on moments greater than the fourth could be calculated, reflecting ever more esoteric properties of the distribution. They rarely are, however. Indeed, even the skewness and kurtosis of distributions are rarely computed. The principal reason is that except for *very* unusual distributions, the bulk of the differences between distributions is captured by statistics based on the first two moments. In general, the amount of extra discriminating information provided by higher moments drops precipitously with each extra moment computed. A second reason is that even where such higher-order differences exist and are pronounced, we are rarely interested in them for the points we wish to make. Most often we are interested in sources of differences in location and variability. Sources of differences in such things as skewness and kurtosis are either mundane and trivial (e.g., ceiling or floor effects—the scores are constrained in some way to an absolute maximum or minimum, such as occurs with test grades, weight, height, etc.) or of no intrinsic interest. We *reduce* the distribution to these few statistics in the belief that we have captured all or most of what is important about the distribution, and discard what remains as “noise”. Ironically, such reduction or throwing away of information is often an important first step in coming to understand the distribution.

*If* the mean and variance capture the bulk of the important differences among distributions in general, then it should also be the case that knowledge of these two statistics should allow us to *recover* at least some aspects of the original distribution. As discussed in the next chapter, two famous mathematical inequalities allow us to do just that.

## 5.2 Questions

1. A distribution of scores with a mean of 40 and a median of 50 would suffer from what kind of skew? Why?
2. A positively-skewed, leptokurtic distribution would appear *how*, relative to *normal* distribution?
3. For the distribution  $X = \{1,2,3,4,5\}$ , compute the skewness and kurtosis. Describe these statistics in words.
4. An experimenter records the time it takes participants to respond to a flash of light by pressing a key, and records a mean time of 120 msec. Such a distribution would likely have what kind of skew? Explain.
5. A *uniform* distribution of scores (i.e., one in which each score is equally likely) would have what kind of kurtosis?

6. For the distribution,  $X = \{2, 2, 3, 3, 3, 4, 4, 4, 4\}$ , compute the skewness and the kurtosis.

## Chapter 6

# Recovering the Distribution

Imagine that you have carefully recorded the mean and standard deviation of your data, but have lost all the raw scores, and you are now confronted with a question such as: Was it possible that more than 20% of the original scores could have been greater than 3 times the mean?<sup>1</sup> Or that, at a minimum, 80% of the scores were within two standard deviations of the mean? Two famous mathematical inequalities provide for (albeit limited) answers to such questions. Although originally derived within probability theory (and famous for that), each can be presented as consequences of simple algebra.

### 6.1 The Markov Inequality

Andrei Andreyevich Markov (or Markoff, 1856–1922) (see Figure 6.1) is responsible for an inequality that allows for the recovery of at least some aspects of the distribution of the numbers from knowledge of the mean. The inequality is restricted to distributions in which all the values are positive and not all are equal to zero, but is otherwise quite general.

The Markov inequality is based on the idea that the mean is the balancing-point of the distribution (see Chapter 2, Section 2.2.2, p. 12). Having a particular mean (and being bounded on the low end by zero) means that the frequencies of scores much greater than the mean are limited. The further the scores are above the mean, the fewer and fewer of them that could occur in the distribution and still be counterbalanced by enough scores lower than the mean (but greater than or equal to zero) to keep the mean at its value. The precise values of these frequencies for any given distribution cannot, of course, be determined from knowledge of the mean alone, but the *maximum* proportions of them can be. The inequality commonly is expressed in two forms, with one a simple arithmetic

---

<sup>1</sup>Careless, yes, but it does happen. More likely, though, is the possibility that you been provided only these (or related statistics) in a research report, say, but wish to know more about the original distribution of scores.



Figure 6.1: Andrei Andreyevich Markov

transform of the other.

Consider first the situation of a distribution,  $X$ , that has  $n$  scores meeting the conditions of all scores greater than or equal to zero, and not all zero. We are interested in what proportion *at most* of these  $n$  scores could be greater than or equal to the value  $c$  (where  $c \geq \bar{X}$ ).<sup>2</sup> Call this proportion  $w$ . Because we are interested in the *maximum* proportion, we need to consider only the situation that would produce the greatest proportion, which occurs when all the scores greater than or equal to  $c$  are all exactly equal to  $c$ , and all the remaining scores are equal to zero, as any other combination would result in less than the maximum proportion. Under these circumstances, the mean can be expressed as the proportion,  $w$ , of scores with the value  $c$  plus the proportion,  $1 - w$ , of scores with the value of zero. The mean,  $\bar{X}$ , then, is just the sum of these two values:

$$\begin{aligned}\bar{X} &= wc + (1 - w)0 \\ &= wc\end{aligned}$$

---

<sup>2</sup>The restriction of  $c \geq \bar{X}$  is simply to ensure that the proportion does not exceed 1.

Re-arranging the equation, the desired maximum proportion,  $w$ , is

$$w = \frac{\bar{X}}{c}$$

That is, for any value of  $c \geq \bar{X}$ , the proportion of scores in the distribution  $X$  greater than or equal to  $c$  is *at most*  $\bar{X}/c$ . That is,

$$p\{X_i \geq c\} \leq \frac{\bar{X}}{c} \quad (6.1)$$

That also means, taking the complementary event, that *at least*  $1 - (\bar{X}/c)$  of the scores are *at most*  $c$ . So, for example, in a distribution having all positive scores (and not all zero) with a mean of 20, *at most* 1/3 of the scores could be greater than or equal to 60 (i.e., 20/60), and *at least* 2/3 must be less than 60. Similarly, *at most* 1/5 (i.e., 20/100) or 20% of the scores could be greater than or equal to 100, and *at least* 80% must be less than 100.

The other common way of expressing the Markov inequality is in terms of *multiples* of the mean of the distribution—that is, where  $c$  ( $c \geq 1$ ) is some multiple of  $\bar{X}$ . If so, then we can re-express equation 6.1 as

$$p\{X_i \geq c\bar{X}\} \leq \frac{\bar{X}}{c\bar{X}}$$

Cancelling  $\bar{X}$  on the right-side of the equation yields:

$$p\{X_i \geq c\bar{X}\} \leq \frac{1}{c} \quad (6.2)$$

This version of the Markov inequality says that for any value of  $c > 1$ , the proportion of scores in the distribution greater than or equal to  $c$  times the mean is *at most*  $1/c$ . That also means, taking the complementary event, that *at least*  $1 - (1/c)$  of the scores are *at most*  $c$  times the mean. Thus, for example, for any distribution whatsoever having all positive scores (and not all zero), *at most* 50% of the scores are greater than or equal to twice the mean, and *at a minimum* 50% of them are less than that. Similarly, *at most* 25% of the scores are greater than or equal to four times the mean, and *at a minimum* 75% of them are less than that. And so on.

Consider the following example in light of the Markov inequality. A student is explaining how he is doing at university. He notes that although the mean grade for his 30 courses is only 50, “. . . that’s because of some very low grades in outrageously-difficult courses like sadistics (*sic*); actually, in more than 21 of these courses my grade has been more than 75!” Given that negative grades are not usually possible, based strictly on the numbers given, is the student’s claim plausible? As  $75 = (1.5)50$  or 1.5 times the mean, Markov’s inequality says that *at most*  $100/1.5 = 67\%$  of the 30 grades, or 20, could be greater than or equal to 75 with a mean of 50. But the student is claiming that more than 21 courses exceeded or equalled 75. So, either the student is misrepresenting the grades, or he has miscalculated their mean.



Figure 6.2: Pafnuti L. Tchebycheff

## 6.2 The Tchebycheff Inequality

Markov was a student of Pafnuti L. Tchebycheff (or Chebychev, 1821–1894) who was responsible for a more general inequality based on both the mean and the variance (see Figure 6.2). Because it uses more information about the distribution than just the mean, it is able to recover even more precise statements about the distribution. Being forced to maintain both a particular mean and variance at their values places further limits on how deviant and with what relative frequency scores can occur. The further constraint provided by the variance also means that the limitation to all positive values can also be lifted.

We can arrive at the Tchebycheff inequality by applying the Markov inequality (equation 6.2) to squared *deviation scores* of the scores of the distribution—the basis for the calculation of variance (see section 3.3), because such scores necessarily meet the requirement of all being greater than zero (for any distribution in which all the scores are not the same). With  $x_i^2$  as a squared deviation score,

and  $\overline{x^2}$  as the mean of such scores, equation 6.2 becomes

$$p\{x_i^2 \geq \overline{cx^2}\} \leq \frac{1}{c}$$

But  $\overline{x^2}$  is just the variance,  $S_X^2$ , of the actual or original  $X$  scores. Furthermore, the proportion of squared deviation scores  $\geq \overline{cx^2}$  is necessarily the same as the proportion of the actual scores that deviate (plus or minus) by  $\sqrt{cS_X^2} = \sqrt{c}S_X$  from the mean,  $\overline{X}$ . Define  $h = \sqrt{c}$ . Then,

$$p\{X_i \geq hS_X\} \leq \frac{1}{h^2} \tag{6.3}$$

Thus, the Tchebycheff inequality states that *in any distribution of scores whatsoever*, for any value of  $h > 1$ , the proportion of scores greater than or equal to  $h$  standard deviations (plus or minus) from the mean is *at most*  $1/h^2$ . Taking the complementary event, that also means that *at least*  $1 - (1/h^2)$  of the scores must lie within  $h$  standard deviations of the mean. So, for example, for any distribution of scores whatsoever, *no more than 25%* of the scores can be beyond 2 standard deviations from the mean, and *at least 75%* of them must lie within 2 standard deviations of the mean. Similarly, *at most 11.11%* of the scores can be beyond 3 standard deviations from the mean, and *at least 88.89%* of them must lie within 3 standard deviations of the mean. And so on.

Consider the following example. A fast-talking, investment promoter is exalting the virtues of a particular investment package. He claims that “On average, people who invest in this package have received a 30% annual return on their investment, and many people have had returns of over 39%.” Intrigued, you ask for, and receive, the information that the average reported is a mean with a standard deviation of 3% for the investment return of 400 people, and that the “many people” referred to “. . . is more than 50”. Based strictly on the numbers given, is there reason to doubt the claims of the promoter? Returns of over 39% are at least 3 standard deviations from the mean. According to the Tchebycheff inequality, with a mean of 30% and a standard deviation of 3%, *at most 11.11%* or 44.44 of the 400 people could have returns more than 3 standard deviations from the mean, but the promoter is claiming that over 50 had returns of 39% or more. The claims don’t add up.

### 6.3 Questions

1. A distribution of 50 scores with a mean of -4 and a standard deviation of 5 can have *at most* how many of those scores greater than 6?
2. A distribution of heights of plants with a mean of 500 and a standard deviation of 200 can have *at most* how many scores greater than 600?
3. An experimenter records the time it takes participants to respond to a flash of light by pressing a key, and records a mean time of 120 msec. *At*

*least* what proportion of the distribution of times must be at most 180 msec (and isn't it amazing that you can know such things)?

4. In the country of Prolifica, the mean number of children per family is 9.
  - (a) Is it possible that 10% of the families have 17 or more children?
  - (b) Is it possible if the standard deviation in number of children per family is 3?

## Chapter 7

# Correlation

Science is generally concerned with whether and the extent to which variables *covary* across cases; in this sense, it is a search for or verification of *structure* or *patterns of relationship* among a set of variables. There are many methods of measuring this *similarity* in variation among variables across cases, but all of them reduce to a basic idea of *redundancy* among a set of variables: the variables go together in some way such that knowledge of one provides some information about another. Tall people tend to be heavier than short people, for example, although there are clear exceptions to this pattern. Height, then, is partially redundant with weight, which is to say that it is *correlated* with weight. Despite its centrality to all of science, a mathematical definition of correlation arrived quite late in the history of science. As with the concepts of evolution by natural selection and the expansion of the universe (and the “big bang”), the concept of correlation is only obvious in retrospect.

### 7.1 Pearson product-moment correlation coefficient

The measure of correlation we’ll focus on is known as the *Pearson product-moment correlation coefficient*, first proposed (in the equation form presented here) by Karl Pearson<sup>1</sup> (1857-1936) in the 1890s (see Figure 7.1), and based on the earlier geometrical insights of Sir Francis Galton (1822-1911), and the work of both F. Y. Edgeworth (1845-1926) and Walter Weldon (1860-1906). This correlation coefficient is denoted by the symbol “r” from earlier work by Galton for “reversion” (and, later, “regression”; see Chapter 8). Pearson derived his

---

<sup>1</sup>Karl Pearson was originally named “Carl” Pearson. Pearson’s Ph.D. was in economics, and he spent years in Germany studying the works of the great German economists, especially Karl Marx. Pearson was so taken with Marx, that, upon returning to England, he officially changed his name to “Karl”. He was one in a long string of “leftist” scholars in the history of statistics, genetics, and evolutionary theory.



Figure 7.1: Karl Pearson

equation for correlation based on bivariate normal spaces, which involves more complicated mathematical techniques than we wish to entertain here. We'll take a different approach, in fact, two of them.

### 7.1.1 Sums of Cross-Products

The first approach exploits a simple mathematical property of sums of cross-products. Consider any two sets of numbers or variables,  $X$  and  $Y$ , derived from the same cases such that  $X_1$  is associated with  $Y_1$ ,  $X_2$  with  $Y_2$ , and, in general,  $X_i$  with  $Y_i$ . If we compute the sum of the cross-products, that is,

$$\sum_{i=1}^n (X_i Y_i)$$

we find that it reaches its maximum value when the order of the values in  $Y$  *matches* as much as possible the order of the values in  $X$ . The sum reaches its minimum value when the values in  $Y$  are ordered as much as possible *backward*

with respect to the order of the values in  $X$ . Intermediate levels of matched ordering result in intermediate sums of cross-products.

As a concrete illustration of this property, let  $X = \{1, 2, 3\}$  and  $Y = \{4, 5, 6\}$ . For this original ordering, the sum of cross-products equals  $1*4 + 2*5 + 3*6 = 32$ . If we compute the sum of cross-products for each of the remaining 5 orderings of the values of  $Y$ ,<sup>2</sup> we get 31 for the orderings  $\{4, 6, 5\}$  and  $\{5, 4, 6\}$ , 29 for orderings  $\{5, 6, 4\}$  and  $\{6, 4, 5\}$ , and a sum of 28 for the last ordering of  $\{6, 5, 4\}$  where the numbers are completely reversed. For these numbers, these sums are perfectly good measures of how well correlated  $X$  is with  $Y$ : if the sum of cross-products equals 32,  $X$  and  $Y$  are as positively correlated as they can be; if the sum is 28, then they are as reversed correlated as they can be; intermediate sums indicate intermediate levels of correlation between these two extremes. Unfortunately, these sums are tied to these specific numbers. Different values for  $X$  and  $Y$  generally will produce different minimum and maximum sums, as will simply using more numbers. The latter problem we can correct, as usual, by computing a *mean* cross-product:

$$\frac{\sum_{i=1}^n (X_i Y_i)}{n}$$

But that doesn't solve the more general problem of different maximum and minimum sums of cross-products for different sets of numbers. Knowing that the mean cross-product for some  $X$  and some  $Y$  is 3.71 tells us little or nothing about the relationship between the two variables unless we also compute the corresponding maximum and minimum mean cross-products for the values in question.

However, if we *standardize* both  $X$  and  $Y$  and compute the mean cross-product of the standardized values, then we'll have a fixed maximum and a fixed minimum, at least theoretically.<sup>3</sup> This mean cross-product of standard scores is the Pearson product moment formula for the correlation coefficient,  $r$ :

$$r_{xy} = \frac{\sum_{i=1}^n Z_{X_i} Z_{Y_i}}{n} \quad (7.1)$$

Because standardizing *centres* the data, computing the mean cross-product of standardized scores also means that positive values of  $r$  are associated with direct or positive relationships, negative values with reversed relationships, and zero with no relationship at all.

<sup>2</sup>We could, of course, just as easily have used all the different orderings or *permutations* of  $X$ , and held the ordering of  $Y$  constant. The result is the same in either case.

<sup>3</sup>*Theoretically* because the maximum and minimum values will only be obtained if the orderings of the standardized values are perfectly matched, *and* the paired deviations are exactly the same in magnitude. Otherwise, even if perfectly matched with respect to order, the computed sum will only approach the theoretical maximum (or minimum).

The theoretical maximum value of  $r$  can be computed as the correlation of a variable with itself (and remembering from section 4.2.1, page 39 that  $\sum_{i=1}^n Z^2 = n$ ):

$$\begin{aligned} \text{MAX}(r_{xy}) &= \frac{\sum_{i=1}^n Z_{X_i} Z_{X_i}}{n} \\ &= \frac{\sum_{i=1}^n Z_{X_i}^2}{n} \\ &= \frac{n}{n} \\ &= 1 \end{aligned}$$

And the theoretical minimum is given as the correlation of a variable with the negative of itself (i.e., each value multiplied by  $-1$ ):

$$\begin{aligned} \text{MIN}(r_{xy}) &= \frac{\sum_{i=1}^n Z_{X_i} (-Z_{X_i})}{n} \\ &= \frac{\sum_{i=1}^n -(Z_{X_i}^2)}{n} \\ &= \frac{-n}{n} \\ &= -1 \end{aligned}$$

Thus,  $-1 \leq r \leq 1$ .

### 7.1.2 Sums of Differences

The other approach characterizes correlation as a measure of the degree to which the paired-values evince similar deviations from their respective means; that is, the extent to which the deviations on one variable are associated with matched deviations on another: large positive deviations on one variable with large positive deviations on the other, large negative deviations on one with large negative deviations on the other, and so on. An inverse or negative correlation would correspond to large positive deviations on one variable associated with large, negative deviations on the other, and so on. One approach to assessing the similarity of the paired deviations would be to compute the average difference between the pairs expressed as standard scores. That is,

$$\frac{\sum_{i=1}^n (Z_{X_i} - Z_{Y_i})}{n}$$

Unfortunately, because  $\sum_{i=1}^n Z_{X_i} = \sum_{i=1}^n Z_{Y_i} = 0$ , this mean is always zero. As usual to eliminate this problem, we can square the paired differences to produce an *average squared difference* or *ASD*:

$$ASD = \frac{\sum_{i=1}^n (Z_{X_i} - Z_{Y_i})^2}{n}$$

#### What values can ASD take?

Again, a perfect positive relationship occurs when  $Z_{X_i} = Z_{Y_i}$ . In that case, obviously,  $ASD = 0$ . On the other hand, the perfect inverse relationship occurs when  $Z_{X_i} = -Z_{Y_i}$ . Substituting that into the *ASD* equation yields:

$$\begin{aligned} ASD &= \frac{\sum_{i=1}^n (Z_{X_i} - (-Z_{X_i}))^2}{n} \\ &= \frac{\sum_{i=1}^n (Z_{X_i} + Z_{X_i})^2}{n} \\ &= \frac{\sum_{i=1}^n (2Z_{X_i})^2}{n} \\ &= \frac{\sum_{i=1}^n 4Z_{X_i}^2}{n} \\ &= 4 \frac{\sum_{i=1}^n Z_{X_i}^2}{n} \\ &= 4 \frac{1}{1} \\ &= 4 \end{aligned}$$

So, a perfect positive relationship yields an *ASD* of 0, and a perfect inverse relationship is given by an *ASD* of 4. Clearly, *no* relationship is given by half of each extreme,  $1/2 * 0 + 1/2 * 4 = 2$ .

As it stands, *ASD* is a perfectly good measure of relationship, but it is confusing to use because:

- no relationship (2) is numerically greater than a perfect positive relationship (0), and
- numerically, a perfect inverse relationship (4) is greater than a perfect positive relationship (0).

We can rectify both problems by subtracting *ASD* from 2. Doing so re-oriens the scale so that inverse relationships are now usefully indicated by negative

numbers (i.e.,  $2 - 4 = -2$ ), positive relationships by positive numbers (i.e.,  $2 - 0 = 2$ ), and no relationship by zero (i.e.,  $2 - 2 = 0$ ). That is,  $-2 \leq 2 - ASD \leq 2$ . But why stay with a range of -2 to +2? If we continue and divide  $(2 - ASD)$  by 2, we obtain the range of  $-1 \leq 1 - \frac{1}{2}ASD \leq +1$  with the values still correctly oriented. This measure is also referred to as  $r$ , and is computed as:

$$\begin{aligned} r_{xy} &= 1 - \frac{1}{2}ASD \\ &= 1 - \frac{\frac{1}{2} \sum_{i=1}^n (Z_{X_i} - Z_{Y_i})^2}{n} \end{aligned} \quad (7.2)$$

It is, in fact, the Pearson product-moment correlation coefficient calculated from differences between pairs of standardized values rather than products. Table 7.1 shows the mathematical equivalence of equation 7.1 with equation 7.2.

## 7.2 Covariance

As noted, the Pearson product-moment correlation coefficient is computed with two variables,  $X$  and  $Y$ , converted to standard-scores so as to limit the maximum and minimum values of  $r_{xy}$  to 1.0 and  $-1.0$ . There is another measure of association based on the same principles that is simply the average cross-product of *deviation scores*, rather than standard scores. It is called the *covariance* of the paired scores, and is defined as follows for the covariance of variables  $X$  and  $Y$ :

$$S_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n} \quad (7.3)$$

Notice how, if  $X$  and  $Y$  were the same variables (i.e.,  $X_i = Y_i$ ), equation 7.3 is simply the definitional formula for variance,  $S_X^2$ . It is used in those situations in which the original measurement scales of the variables are considered to be an important part of their relationship.

## 7.3 Other Correlational Techniques

The Pearson product-moment correlation coefficient can be computed for paired values of any two variables; regardless of what the variable values actually refer to (e.g., continuous measurements, dichotomous or polytomous categories, ranks, or any combination), the  $r$ -value calculated will reveal the correlation (i.e., mean cross-product or mean squared difference of standard scores) of the paired-values as recorded, whether otherwise meaningful or not. Some combinations of types of variables, however, allow for simpler calculation of  $r$ . Although these different techniques for calculating  $r$  typically are given different names, it should be remembered that the resulting value *is* the Pearson product-moment correlation coefficient, and *is* the exact same value as would be obtained had either equation 7.1 or equation 7.2 been applied to the same data.

$$\begin{aligned}
r_{xy} &= 1 - \frac{1}{2} ASD \\
&= 1 - \frac{1}{2} \frac{\sum_{i=1}^n (Z_{X_i} - Z_{Y_i})^2}{n} \\
&= 1 - \frac{1}{2n} \sum_{i=1}^n (Z_{X_i}^2 - 2Z_{X_i}Z_{Y_i} + Z_{Y_i}^2) \\
&= 1 - \frac{1}{2n} \left( n - 2 \sum_{i=1}^n Z_{X_i}Z_{Y_i} + n \right) \\
&= 1 - \frac{1}{2n} \left( 2n - 2 \sum_{i=1}^n Z_{X_i}Z_{Y_i} \right) \\
&= 1 - \left( 1 - \frac{\sum_{i=1}^n Z_{X_i}Z_{Y_i}}{n} \right) \\
&= 0 + \frac{\sum_{i=1}^n Z_{X_i}Z_{Y_i}}{n} \\
&= \frac{\sum_{i=1}^n Z_{X_i}Z_{Y_i}}{n}
\end{aligned}$$

Table 7.1: Demonstration that the average squared difference formula for  $r$  and the cross-products formula for  $r$  are equivalent.

Car	Sandy (S)	Chris (C)	$Z_S$	$Z_C$	$Z_S Z_C$
Topaz	1	3	-1.42	0	0
Geo	3	1	0	-1.42	0
Volvo	2	4	-0.71	0.71	-0.5
Civic	5	5	1.42	1.42	2.02
Sprint	4	2	0.71	-0.71	-0.5
mean	3	3	0	0	0.2

Table 7.2: Demonstration that the Spearman formula is equivalent to the Pearson cross-product formula applied to the data in ranks (equation 7.1).

### 7.3.1 The Spearman Rank-Order Correlation Coefficient

When the data for both variables are in ranks, either because the data were collected that way (e.g., two judges are asked to rank-order their preferences for 20 automobiles) or because the data were converted to ranks (e.g., because only the ordinal properties of the variables were of interest), the unique properties of ranks may be exploited to provide a simple or “short-cut” method for computing the correlation.

When data for both variables are in ranks, we know *a priori* that the minimum value is 1, and the maximum value is  $n$ . We also know from section 7.1.1 that the sum of cross-products of the paired ranks reaches its maximum when rank 1 on one variable is associated with rank 1 on the other variable, 2 with 2, and so on. Furthermore, because the data are in ranks, we know that this maximum sum of cross-products depends only on  $n$ . Given these constraints, it turns out<sup>4</sup> that the correlation between the paired ranks can be computed as follows, using  $D_i$  to indicate the difference between the rank of the  $i$ th value on one variable and its corresponding rank on the other variable:

$$r = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)} \quad (7.4)$$

For example, suppose the ranked preference of five cars (e.g., a Topaz, a Geo, a Volvo 244, a Civic, and a Sprint, respectively) from Sandy were  $S = \{1, 3, 2, 5, 4\}$ ; that is, Sandy preferred the Topaz most and the Civic least. Suppose further that Chris ranked the same 5 cars as  $C = \{3, 1, 4, 5, 2\}$ ; that is, Chris preferred the Geo most and Civic least. As shown in Table 7.2, the Spearman rank-order correlation coefficient for these ranks would be:  $1 - 6 * (2^2 + 2^2 + 2^2 + 0 + 2^2) / (5^3 - 5) = 1 - (6 * 16) / 120 = 1 - .8 = .2$ .

<sup>4</sup>The precise derivation is left as an exercise for the reader. In doing so, note that the formula is based on *squared differences* rather than cross-products. Accordingly, as in section 7.1.2, the *minimum* sum should be associated with the pairs precisely matched in rank, and the maximum sum when the pairs are inversely matched; hence, the subtraction from 1 in equation 7.4.

### Tied Ranks

Sometimes, the ranking procedure results in *tied ranks*. It might be the case, for example, that the individual or process doing the ranking can't discriminate between two or more cases; they all should be ranked "3", for instance. Or, it may be the case that the data to be ranked contain a number of equal values. Tied ranks are problematic because equation 7.4 assumes that each value has a unique rank. There are two ways of handling this situation. The tied cases could simply be ignored, and the correlation computed for the remaining cases, or the tied ranks could be adjusted. In the latter case, the ranks that would normally be associated with the values are averaged, and the average used as the rank for each of the values. For example, if the ranks in question were 7, 8, and 9, the ranks would be averaged to 8, and the rank of 8 assigned to each of the three values. Applying equation 7.4 to ranks adjusted in this way tends to *overestimate* the absolute value of  $r$  that would be obtained with untied ranks, but as long as the number of tied ranks is small, the effect is usually trivially small.

### 7.3.2 The Point-Biserial Correlation Coefficient

Suppose you were interested in the relationship between scores on some continuous measure (e.g., height, weight, proportion recalled, etc.) and some dichotomous categories of cases that produced the scores (e.g., male vs. female, young vs. old, Capulets vs. Montagues, experimental group vs. control group, etc.). You could compute the correlation between these two variables by assigning one number (say 0) to every case in one category, and any other number (say 1) to every case in the alternate category, and then apply the equation for the Pearson product-moment correlation coefficient to the data as transformed. Because the correlation coefficient is based on standardized scores, it doesn't matter which two numbers you choose (except for the *sign* of the coefficient), as long as they are used consistently.

Because the precise values of numbers used for the dichotomous categories don't matter, all that is important about this variable for determining the relationship between the two variables is the *proportions* of cases in each of the two categories. Similarly, as any variation of the continuous variable *within* each of the groups is just noise (i.e., variation *unrelated* to the categorical variable) with respect to the relationship between the variables, what is important about the continuous variable are the *means* of the two groups, and in particular the ratio of the difference between the two means and the variability among cases as a whole. Accordingly, a simpler formula for computing  $r$  is available in these circumstances, and it is called the *point-biserial correlation coefficient*. Using  $X$  as the dichotomous variable, with  $p$  and  $q$  as the proportions in the two categories, and  $Y$  as the continuous variable, with  $\bar{Y}_0$  as the mean for group 0,  $\bar{Y}_1$  as the mean for group 1, and  $S_Y$  as the standard deviation of variable  $Y$ ,

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{S_Y} \sqrt{pq} \quad (7.5)$$

Group	Group Code	Errors	$Z_G$	$Z_E$	$Z_G Z_E$
control	0	1	-1.22	-1.5	1.83
control	0	4	-1.22	0	0
experimental	1	3	0.82	-0.5	-0.41
experimental	1	5	0.82	0.5	0.41
experimental	1	7	0.82	1.5	1.23
mean	0.6	4			0.61
std. dev.	0.49	2			

Table 7.3: Demonstration that the point-biserial formula is equivalent to applying the Pearson cross-product formula to the data for which the categorical variable is given a binary code (equation 7.1).

As an example, suppose 2 people (the control group) were given ample time to complete a task consisting of 10 problems, and scored  $\{1, 4\}$  errors, respectively. Suppose a further 3 people (the experimental group) were given the same problems, but were forced to work on them under a severe time deadline, and scored  $\{3, 5, 7\}$  errors, respectively. The mean number of errors for the control group is 2.5, and they constitute  $2/5$  or  $.4$  of the participants; the mean number of errors for the experimental group is 5, and they constitute  $3/5$  or  $.6$  of the participants; the standard deviation of errors for the 5 people as a whole is 2. Hence, the correlation between number of errors and group membership is  $r = (5 - 2.5)/2 * \sqrt{.4 * .6} = 1.25 * 0.49 = 0.61$ . That is, there would be a moderately high correlation between the the number of errors produced and whether or not the task was done under a severe deadline. The same value would have been obtained by assigning a zero as their  $X$ -values to each of the 2 cases in the control group, and a 1 as their  $X$ -values to each of the 3 cases in the experimental group, and then applying either of the equations for the Pearson product-moment correlation coefficient from section 7.1, as shown in Table 7.3.

The point-biserial correlation coefficient is mathematically related to another statistic, known as the  $t$ -statistic. See Chapter 14.

### 7.3.3 And Yet Other Correlational Techniques

There are still other techniques that are just variant ways of computing the Pearson product-moment correlation coefficient taking computational advantage of various constraints imposed by the data. For example, the  $\phi$  coefficient is a simplified way of computing the Pearson  $r$  when *both* variables are dichotomous. It is closely related to another statistic known as *chi-square*, and is better discussed in that context. Still other coefficients, although not mathematically equivalent to Pearson  $r$ , can be seen as *generalizations* of the general principle for such situations as the correlation between a continuous variable and a polytomous one, or the correlation between two polytomous variables. The former is an example of what is known as *multiple correlation*, and the latter is an example of *categorical association*. Both are discussed in subsequent chapters.

## 7.4 Questions

- Each of you and a friend rank the 5 recently-released, currently-playing movies, with 1 as the most preferred and 5 as the least preferred. The results are as follows:

Movie	You	Friend
A	3	1
B	4	2
C	1	5
D	2	4
E	5	3

How well are your ratings of the movies related to those of your friend?

- Your statistically unsophisticated coworker has asked you to take on a task for the gym where she works. She wants to be able to claim that “Smarter people choose Jim’s Gym.” You suggest that she correlate gym membership with IQ. She collected the following information from 10 people:

Person	Gym	IQ
Arnold	Tim’s	120
Hans	Tim’s	110
Franz	Tim’s	100
Lou	Tim’s	80
Charles	Tim’s	90
PeeWee	Jim’s	90
Percival	Jim’s	90
Les	Jim’s	110
Dilbert	Jim’s	140
Edwin	Jim’s	120

What is the correlation between gym membership with IQ? Does it appear to be the case (at least for these data) that “Smarter people choose Jim’s Gym.”?

- On recent sadistics (sic) test, the 16 psychology majors in the class achieved a mean grade of 80, and the 4 business majors in the class achieved a mean grade of 76. If the standard deviation for the 20 students was 2, what was the correlation between major and sadistics (sic) grade?
- An innvestigator measures the length and girth (around the midsection) of 500 gila (“heela”) monsters. Converting each of these measures to standard scores, she finds the sum across gila monsters of the squared difference between length and girth is equal to 100. What is the Pearson product-moment correlation coefficient relating gila length to girth? Describe the relationship in words.

5. Below are the results of two horse races involving the same 5 horses:

Horse	Race 1	Race 2
Never Fails	3	3
Sadly Happy	1	2
Mom's Apple	2	1
Ski Bunny	4	5
Tawanda	5	4

What is the correlation between the two races? Describe the relationship in words.

## Chapter 8

# Linear Regression

Finding that two variables covary is also to say that from knowledge of one the other may be *predicted*, at least to some degree. If this covariation is captured as a correlation coefficient, then the *degree* or extent to which one variable can be predicted by the other is given by its value. But exactly what kind of prediction is it? The Pearson product-moment correlation coefficient (and its algebraic variants) is a measure of the *linear intensity* of the relationship between two variables; it measures the extent to which the values of one variable may be described as a *linear transformation* of the values of another variable. In fact, it was from the observations that one variable could be seen as an approximate linear function of another that led to the development of the concept of the correlation coefficient in the first place.

### 8.1 Linear Regression

In his classic paper of 1885, *Regression toward mediocrity in hereditary stature*, Galton noted that in predicting the heights of children from the heights of their parents, tall parents tended to produce tall children, but generally *less* tall than themselves. Similarly, short parents tended to produce short children, but generally *less* short than themselves. Thus, in *predicting* the heights of children from the heights of their parents, the most accurate predictions were obtained, on average, by predicting heights in the same direction as, but less extreme than that of the parents. He noted that if one plotted the *median* (later, following Karl Pearson, *mean*) height of the children as a function of the height category to which their fathers belonged, an approximately straight line could be drawn through the points, but that the *slope* of this line was generally some fraction less than 1.0. A slope less than 1.0 means that any prediction based on the line of the children's height from that of their parents would be less extreme because it would be a predicted fraction of that of the parents.

Based on these and other data, Galton theorized that physical traits such

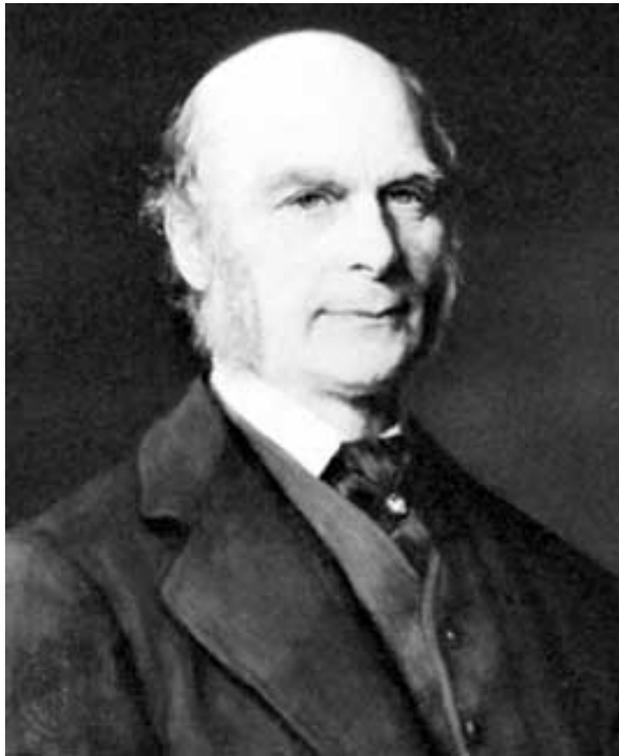


Figure 8.1: Sir Francis Galton

as height *regressed* toward the mean height over generations. Accordingly, he labeled the line the *regression line*. The method of prediction that he developed thus came to be known as *linear regression*, the unfortunate name it retains to this day.<sup>1</sup> Galton recognised that the slope of the regression line—the scaling or *regression coefficient* for the prediction—was an important measure of the relationship (or “co-relation”—Galton’s term) between two variables, but other than his graphical or geometrical approach, he did not develop a simple algebraic method for its computation. Followers of Galton, notably Edgeworth, Weldon, and ultimately Pearson worked on the problem and, as discussed in the chapter on correlation, solved it in the form of the Pearson Product-Moment Correlation Coefficient—the value this linear regression coefficient or slope takes when the

---

<sup>1</sup>Unfortunate because the phenomenon Galton observed has nothing directly to do with either biology or inter-generational “regression”. Rather, it is a function simply of less than perfect correlation between the two variables: the poorer the correlation, the greater the “regression” of predictions toward the mean. For example, the same regression phenomenon is found when predicting *parental* height from that of their children: tall children, for example, tend to have tall parents but not, on average, as tall as themselves. Incidentally, Galton was well aware that the phenomenon was not restricted to biological factors over generations, but for many years this misinterpretation persisted.

variables are *regressed* (as we now say) on one another in standard score form.

### 8.1.1 The Regression Equation: $Z'_{Y_i} = r_{xy}Z_{X_i}$

Consider the case where there is only a *moderate* correlation between variables  $X$  and  $Y$ , say,  $r_{xy} = .5$ . For this to be true,  $\sum_{i=1}^n Z_{X_i}Z_{Y_i}$  must equal  $\frac{1}{2}n$ .<sup>2</sup> How could this occur? If each  $Z_Y$  of every  $Z_XZ_Y$  pair were equal to  $\frac{1}{2}Z_X$  then

$$\begin{aligned} r_{xy} &= \frac{\sum_{i=1}^n Z_{X_i}Z_{Y_i}}{n} \\ &= \frac{\sum_{i=1}^n Z_{X_i} \cdot .5Z_{X_i}}{n} \\ &= .5 \frac{\sum_{i=1}^n Z_{X_i}Z_{X_i}}{n} \\ &= .5 \frac{\sum_{i=1}^n Z_{X_i}^2}{n} \\ &= .5 \frac{n}{n} \\ &= .5 \end{aligned}$$

But it is not necessary that each  $Z_{Y_i} = .5Z_{X_i}$  for this statement to be true, rather that *on average* they do so. That is,  $r_{xy} = .5$  merely indicates that, *on average*,  $Z_Y = .5Z_X$  (or  $Z_X = .5Z_Y$ ). If  $r_{xy} = .5$ , then, what is the *best* prediction of  $Z_Y$  given  $Z_X$ ? Given the foregoing, it would appear that  $Z_Y = .5Z_X$  would be the best prediction. How is this the *best* prediction? For any particular  $Z_XZ_Y$  pair, it may not be. Using  $Z'_{Y_i}$  as the *predicted* value of  $Z_{Y_i}$  from  $Z_{X_i}$ , we may find that the difference,  $Z_{Y_i} - Z'_{Y_i}$ , is quite large for some pairs. But we know that the sum of squared deviations from the mean is a minimum in that any value other than the mean will produce a larger sum of squared deviations from it. So, predicting the *average*  $Z_Y$  for a given  $Z_X$  will produce the smallest *average* squared difference between the predicted and the actual values. That is, *on average*, predicting the average  $Z_Y$  for a given  $Z_X$  produces the smallest *error of prediction*, at least as measured in terms of this *least-squares criterion*.<sup>3</sup> But, as just noted, the *average*  $Z_Y$  given  $Z_X$  is equal to  $r_{xy}Z_X$ . That is, using  $Z'_{Y_i}$

<sup>2</sup>It will be remembered that  $r_{xy} = \frac{\sum_{i=1}^n Z_{X_i}Z_{Y_i}}{n}$ , so  $\sum_{i=1}^n Z_{X_i}Z_{Y_i} = r_{xy}n$ .

<sup>3</sup>One can start with this notion of minimising the sum of squared deviations of the actual from the predicted values, and ask what linear expression minimises it using differential calculus. The result is the same in either case.

as the predicted value, the linear equation

$$Z'_{Y_i} = r_{xy}Z_{X_i} \quad (8.1)$$

produces the lowest error of prediction (in terms of least-squares) in  $Z_Y$  for all  $r_{xy}$ . Similarly,

$$Z'_{X_i} = r_{xy}Z_{Y_i} \quad (8.2)$$

produces the lowest error of prediction (in terms of least-squares) in  $Z_X$  for all  $r_{xy}$ . Note that unless  $r_{xy}$  equals 1 or -1, the predicted value of  $Z'_{Y_i}$  is always less extreme or deviant than the  $Z_{X_i}$  value from which the prediction is made. This reduction in deviance in predicted values is Galton's "regression", and occurs strictly as a function of imperfect correlation: the poorer the correlation between  $X$  and  $Y$ , the greater the regression in the predicted values.

### 8.1.2 From standard scores to raw scores

It is not necessary to convert the data to standard scores in order to compute a regression equation. As we know how to move from raw-scores to standard scores and back (i.e.,  $X_i = Z_i S_X + \bar{X}$ ), the simple standard score regression equation can be manipulated algebraically to produce a regression equation for for the data in raw-score form:

$$\begin{aligned} Y'_i &= Z'_{Y_i} S_Y + \bar{Y} \\ &= r_{xy} Z_{X_i} S_Y + \bar{Y} \\ &= r_{xy} \frac{X_i - \bar{X}}{S_X} S_Y + \bar{Y} \\ &= r_{xy} \frac{S_Y}{S_X} X_i + \bar{Y} - r_{xy} \frac{S_Y}{S_X} \bar{X} \end{aligned} \quad (8.3)$$

Despite its formidable-looking form, equation 8.3 still expresses the predicted value,  $Y'_i$ , as a linear function of  $X_i$ , with  $r_{xy} \frac{S_Y}{S_X}$  as the slope, and  $\bar{Y} - r_{xy} \frac{S_Y}{S_X} \bar{X}$  as the intercept. A similar regression equation for predicting raw-score  $X$ -values from raw-score  $Y$ -values is given by:

$$X'_i = r_{xy} \frac{S_X}{S_Y} Y_i + \bar{X} - r_{xy} \frac{S_X}{S_Y} \bar{Y} \quad (8.4)$$

Note that unless the standard deviations are equal,  $\bar{Y} = \bar{X}$ , and  $r_{xy} = \pm 1$ , the two regression lines are *not* the same.<sup>4</sup>

As an example, consider the hypothetical data shown in Table 8.1 representing the heights in inches of five adult women paired with that of their closest (in age) adult brothers. As shown, the correlation in height between sister and brother is

<sup>4</sup>There is a line, called the *first principal component*, that is the average of the two regression lines. Whereas the regression line of  $Y$  on  $X$  minimizes  $\sum(Y - Y')^2$ , and the regression of  $X$  on  $Y$  minimizes  $\sum(X - X')^2$ , the principal component line minimizes the sum of squared deviations between obtained and predicted along both axes simultaneously.

Family	Actual Height		Standard Score		$Z_s Z_b$
	Sister	Brother	$Z_s$	$Z_b$	
Smith	54	70	-1.5	-0.5	0.75
Jones	60	66	-0.5	-1.5	0.75
Young	63	72	0.0	0.0	0.00
Wong	66	78	0.5	1.5	0.75
Chang	72	74	1.5	0.5	0.75
Mean	63	72	0	0	$0.6 = r_{sb}$
S.D.	6	4	1	1	

Table 8.1: Example regression data representing the heights in inches of 5 women and their brothers expressed as both raw and standard scores. Column  $Z_s Z_b$  is the cross-product of the paired standard ( $Z$ -) scores.

$r_{sb} = 0.6$  — a moderate, positive correlation. Figures 8.2 and 8.3 plot these data in raw and standard score form, respectively. Superimposed on each figure are the regression lines of brother’s height regressed on sister’s, and sister’s height regressed on brother’s.

### 8.1.3 Correlation and Regression: $r_{xy} = r_{y'y}$

The regression equation expresses the predicted values of  $Y$ ,  $Y'$ , as a *linear function* of the obtained  $X$ -values, with  $r_{xy}$  as the slope of this line when the scores are expressed as standard scores. As standard scores are themselves linear functions of the raw scores, the correlation between  $Y$  and *any* linear transformation of  $X$  is also equal to  $r_{xy}$ . Consequently, as  $Y'$  is a linear transformation of  $X$ , the correlation between  $X$  and  $Y$  is also the correlation between  $Y$  and  $Y'$ . Said differently, the correlation,  $r_{xy}$ , is the correlation between  $Y$  and the linear transformation of  $X$  that best (in the least-squares sense) matches the values of  $Y$ ;  $r_{xy} = r_{y'y}$ . So, if  $r_{xy}$  (or, equivalently,  $r_{y'y}$ ) is low that means that predicted  $Y'$  scores are a poor match to the actual  $Y$ -values; whereas, if  $r_{xy}$  is high, the predicted  $Y'$  scores are a good match to the actual  $Y$ -values.

### 8.1.4 The standard error of estimate

Earlier, we introduced the notion of “best-fitting line” in the sense that it minimises the sum of squared deviations of the actual from their predicted values. This sum,

$$SS_{res} = \sum_{i=1}^n (Y_i - Y'_i)^2$$

is referred to as the *residual sum-of-squares* or  $SS_{res}$  because it represents the residual or what is left over *after* taking the linear prediction from variable  $X$

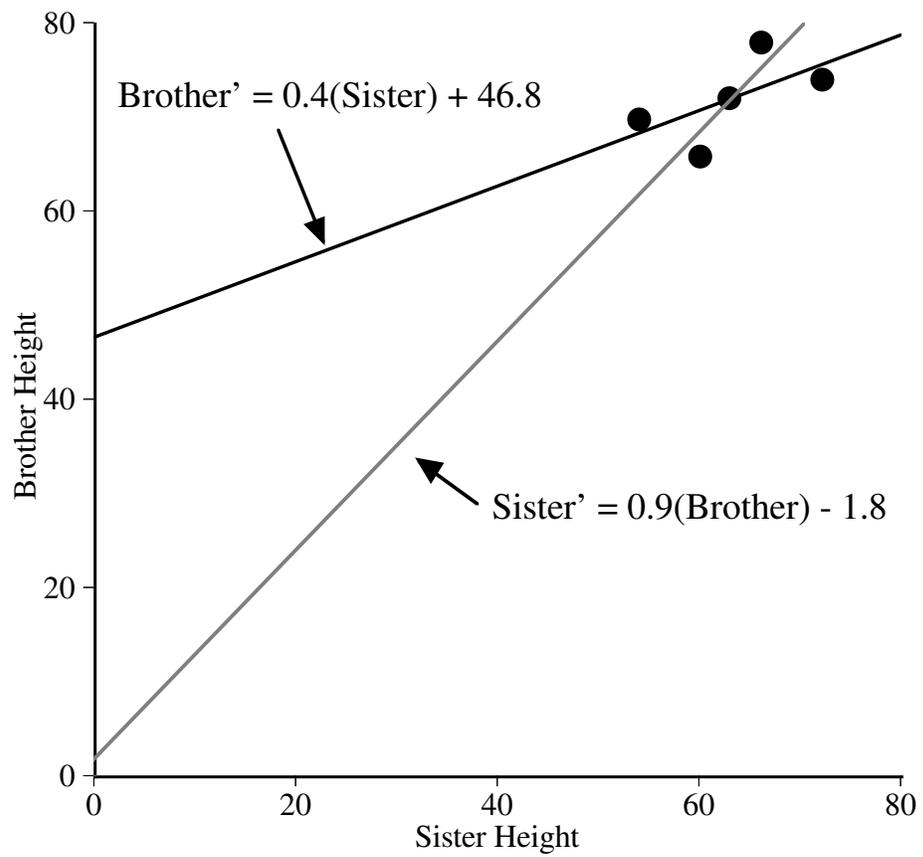


Figure 8.2: Scatterplot and regression lines for the hypothetical data from Table 8.1.

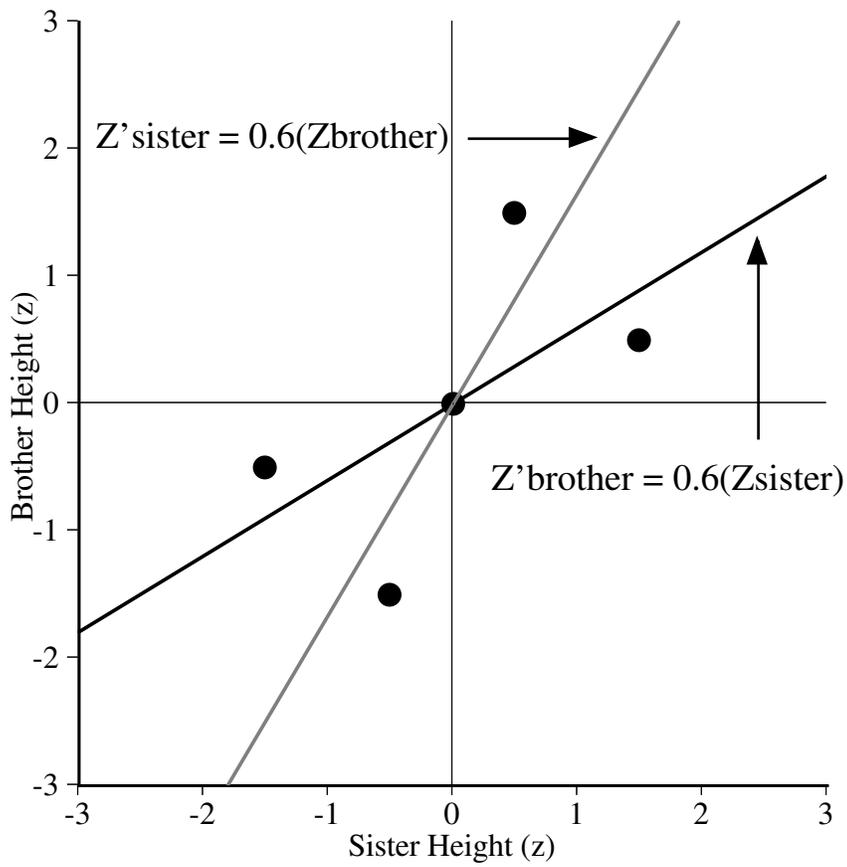


Figure 8.3: Scatterplot and regression lines for the standardized hypothetical data from Table 8.1.

into account. Expressed as an average or *mean-square residual*,  $MS_{res}$ ,

$$MS_{res} = \frac{\sum_{i=1}^n (Y_i - Y'_i)^2}{n}$$

yields a *variance* measure of the deviation of the actual from their predicted values—the variance of the errors of prediction. The positive square-root of  $MS_{res}$  provides a measure in terms of the original units of the data, and is known as the *standard error of estimate*,  $S_{Y-Y'}$ :

$$S_{Y-Y'} = \sqrt{\frac{\sum_{i=1}^n (Y_i - Y'_i)^2}{n}} \quad (8.5)$$

It is usually more convenient to compute the standard error of estimate with the following formula, which can be shown to be equivalent by algebraic manipulation:

$$S_{Y-Y'} = S_Y \sqrt{1 - r_{xy}^2} \quad (8.6)$$

Clearly, if  $r_{xy} = 0$ , the *best* (in the least-squares sense) prediction of every  $Y_i$  from  $X_i$  is simply the mean of  $Y$ ,  $\bar{Y}$ , and, as equations 8.5 and 8.6 indicate, the error of prediction in this case is just the standard deviation of  $Y$ . On the other hand, if  $r_{xy} = \pm 1.0$ , then the standard error of estimate is zero. The standard error of estimate can be thought as capturing the variability of the  $Y$ -values for a particular  $X$ -value, averaged over all values of  $X$ . Low values of  $S_{Y-Y'}$  indicate that the average spread of the actual values of  $Y$  around the value predicted from  $X$  is also low—that, on average, the predicted values are quite accurate. Conversely, values of  $S_{Y-Y'}$  close to the standard deviation of  $Y$  indicate that predictions from  $X$  are little better than simply predicting the mean of  $Y$  in every case—that knowledge of  $X$  contributes little to the (linear) prediction of  $Y$ .

The standard error of estimate also provides another interpretation of the correlation coefficient. It can be shown by algebraic manipulation of equation 8.6 that:

$$\begin{aligned} r_{xy}^2 &= 1 - \frac{S_{Y-Y'}^2}{S_Y^2} \\ &= \frac{S_Y^2 - S_{Y-Y'}^2}{S_Y^2} \end{aligned}$$

That is, the squared correlation coefficient,  $r_{xy}^2$ , represents the reduction in the proportion of the original variance of  $Y$  that is provided by knowledge of  $X$ . The better or more accurately  $Y$  is (linearly) predicted by  $X$ , the greater the value of  $r_{xy}^2$  and the greater the proportion of the variance of  $Y$  “accounted for” by

knowledge of  $X$ . For this reason, the square of the correlation coefficient is called the *coefficient of determination*. This proportion is, in fact, that proportion of the original variance of  $Y$  that corresponds to the variability of the predicted values of  $Y$  (i.e., the variability of the points along the regression line):

$$\begin{aligned} S_{Y'}^2 &= \frac{\sum_{i=1}^n (Y'_i - \bar{Y})^2}{n} \\ &= S_Y^2 r_{xy}^2 \end{aligned}$$

Hence, what has been accomplished by the regression is the *partitioning* of the *total sum-of-squares*,  $SS_{tot}$ , into two pieces, that which can be “accounted for” by the regression,  $SS_{reg}$ , and that which cannot,  $SS_{res}$ :

$$\begin{aligned} SS_{tot} &= SS_{reg} + SS_{res} \\ \sum_{i=1}^n (Y - \bar{Y})^2 &= \sum_{i=1}^n (Y' - \bar{Y})^2 + \sum_{i=1}^n (Y - Y')^2 \end{aligned}$$

The term  $1 - r_{xy}^2$  is called the *coefficient of non-determination* because it represents the proportion of the original variance of  $Y$  that is “unaccounted for” or not determined by  $X$ —that is, not linearly related to  $X$ . The positive square-root of the coefficient of non-determination, as in equation 8.6, is called the *coefficient of alienation* because it reflects the “alienation” of the  $Y$  scores from the  $X$  scores in unsquared terms (as with  $r_{xy}$ ).

### 8.1.5 The Proportional Increase in Prediction: PIP

The coefficient of alienation plays a role in yet another interpretation of the correlation coefficient. One of the difficulties in interpreting the correlation coefficient is that it is not immediately obvious how good or useful a particular correlation value is, or how much better one value is relative to another. Although there is no question that a correlation 0.5, for example, is more useful than one of 0.3, just how much more useful is unclear.

A good measure of “usefulness” is predictability. In terms of prediction, what is important about the correlation coefficient is the degree to which it improves the prediction of the values on one variable from knowledge of values on another relative to no such knowledge. As we’ve seen, a large correlation reduces the *error of estimate* and thereby increases the accuracy of prediction to a greater extent than does a small or non-existent correlation. How much a particular correlation improves the prediction (or reduces the error of estimate) is provided by a variant of the coefficient of alienation, called the *proportional increase in prediction* or *PIP*:

$$PIP = 1 - \sqrt{1 - r_{xy}^2}$$

$ r_{xy} $	$\sqrt{1 - r_{xy}^2}$	$1 - \sqrt{1 - r_{xy}^2}$
0.000	1.000	0.000
0.100	0.995	0.005
0.200	0.980	0.020
0.300	0.954	0.046
0.400	0.917	0.083
0.500	0.866	0.134
0.600	0.800	0.200
0.700	0.714	0.286
0.800	0.600	0.400
0.866	0.500	0.500
0.900	0.436	0.564
0.950	0.312	0.688
0.990	0.141	0.859
0.999	0.045	0.955

Table 8.2: Proportional increase in prediction as a function of selected values of  $|r_{xy}|$ . Column 2 is the coefficient of alienation, and column 3 is the proportional increase in prediction or PIP relative to a correlation of zero.

Table 8.2 reports the values of both the coefficient of alienation and PIP for selected values of  $|r_{xy}|$ . Note that compared to a correlation of zero, a correlation of 0.6 produces an error of estimate that is 0.8 or 80% of what it would be (a reduction of 20% in error of estimate or a 20% improvement in prediction), whereas a correlation of 0.3 produces an error of estimate that is only 0.95 or 95% of what it would be with a correlation of zero (i.e., only a 5% reduction in error of estimate or a 5% improvement in prediction). The correlation must be as high as 0.866 before the error of estimate is reduced by one-half (or the prediction improved by 50%) relative to a correlation of zero. Also note that the difference in improvement of prediction between a correlation of 0.7 and a correlation of 0.9 is approximately the same as that between 0.2 and 0.7. That is, not only does the prediction improve with larger correlations, but the *rate* of improvement also increases, as shown in Figure 8.4.

## 8.2 Questions

- The correlation between two tests, A and B, is found to be  $r = .6$ . The standard deviation of test A is 10 and that of test B is 5. If every score of test A is predicted from the corresponding test B scores by virtue of the correlation between them, what would be the standard deviation of
  - the *predicted scores* (i.e.,  $S_{Y'}$ )?
  - the *residuals* (i.e.,  $S_{Y-Y'}$ )?

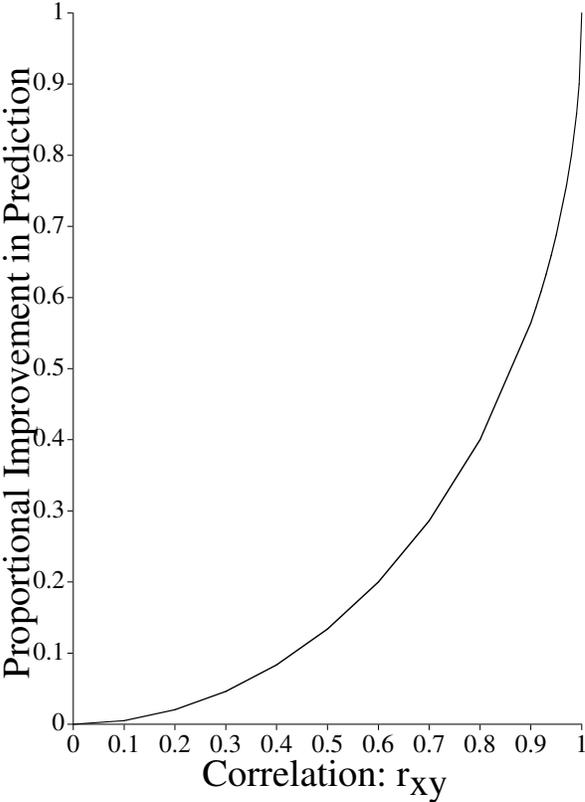


Figure 8.4: Proportional increase in prediction as a function of  $|r_{xy}|$ .

2. You have developed a test to predict performance on stats midterms and hope to market this test to students who prefer drinking beer to studying for midterms (this could be lucrative). You give the test to 5 students before their first midterm, wait a few days, then obtain their midterm scores. Here are your data:

Student	Predictive Test	Midterm
A	5	75
K	6	70
V	7	85
J	4	80
C	3	65

- (a) What is the regression equation to predict the midterm score from your predictive test?
- (b) How much of the variability in test scores does your predictive test account for?
3. Catbert, the evil human resources director, complains that the Wesayso corporation for which he works fosters mediocrity. His evidence for this claim is that people whose first performance review upon entering the company is very poor typically improve on subsequent reviews whereas people whose first performance review is stellar typically get worse on subsequent reviews. Explain to Catbert why his conclusion doesn't follow from his "evidence".
4. Explain why the  $y$  intercept of the regression line for a set of  $Z$  (standard) scores is always zero.
5. With  $r_{xy} = .8$ ,  $S_y = 10$ , and  $S_x = 5$ , if  $Y$  scores are predicted from the  $X$  scores on the basis of the correlation between them, what would be the standard deviation of the *predicted scores*?
6. Research has shown that the best students in their first year of university on average do less well at the end of their last year. However, the very worst students in first year tend on average to do better at the end of their last year. Allen Scott (a.k.a. the class n'er-do-well) argues that these data prove that university hurts the best students, but helps the worst students. Explain to poor Allen why his conclusion doesn't necessarily follow. What assumptions is he making?
7. You collect the following grades in Advanced Egg Admiration (AEA) and Basket-weaving (BW) for five students who have taken both courses:

Student	AEA	BW
Sandy	54	84
Sam	66	68
Lauren	72	76
Pat	60	92
Chris	63	80

Based on Sandy's BW grade, and the correlation between AEA and BW grades, what is Sandy's predicted grade in AEA?



## Chapter 9

# Partial and Multiple Correlation

### 9.1 Partial Correlation

Consider the situation of *three* intercorrelated variables  $X$ ,  $Y$ , and  $Z$ , such that  $r_{xy}$  is the correlation between  $X$  and  $Y$ ,  $r_{xz}$  is the correlation between  $X$  and  $Z$ , and  $r_{yz}$  is the correlation between  $Y$  and  $Z$ . Suppose the interest is in the correlation between  $X$  and  $Y$  in the *absence* of their correlations with  $Z$ . Expressed in terms of standard scores, we can produce the residual  $Z_X$  and  $Z_Y$  scores from their respective regression equations with  $Z$ . That is, the residual  $Z_X$  scores—the residuals from the values *predicted* from  $Z_Z$ —are given by  $Z_X - Z'_X = Z_X - r_{xz}Z_Z$ . Similarly, the residual  $Z_Y$  scores are given by  $Z_Y - Z'_Y = Z_Y - r_{yz}Z_Z$ . By definition, these residual scores are uncorrelated with  $Z$ . Hence, the correlation between these residual standard scores represents the correlation between  $X$  and  $Y$  in the absence of any correlation of either variable with  $Z$ .

The standard deviation of these residual standard scores is equal to  $\sqrt{1 - r_{xz}^2}$  and  $\sqrt{1 - r_{yz}^2}$ , respectively (see section 8.1.4). So, dividing each of these residuals by their respective standard deviations converts them into standardized residuals, or standard scores in their own right.

#### 9.1.1 Part or Partial correlation

As such, the correlation between them can be computed as the average cross-product of paired standard scores (see Chapter 7 on correlation), yielding the following expression for the *partial correlation* between variables  $X$  and  $Y$  in the

$$\begin{aligned}
r_{(xy).z} &= \frac{1}{n} \left[ \frac{\sum (Z_X - r_{xz}Z_Z)(Z_Y - r_{yz}Z_Z)}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}} \right] \\
&= \frac{1}{n\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}} \sum [Z_X Z_Y - Z_X r_{yz} Z_Z - Z_Y r_{xz} Z_Z + Z_Z^2 r_{xz} r_{yz}] \\
&= \frac{1}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}} \left[ \frac{\sum Z_X Z_Y}{n} - \frac{\sum Z_X Z_Z}{n} r_{yz} - \frac{\sum Z_Y Z_Z}{n} r_{xz} + \frac{\sum Z_Z^2}{n} r_{xz} r_{yz} \right] \\
&= \frac{1}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}} [r_{xy} - r_{xz} r_{yz} - r_{yz} r_{xz} + r_{xz} r_{yz}] \\
&= \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}}
\end{aligned}$$

Table 9.1: Derivation of the expression for the partial correlation between variables  $X$  and  $Y$  controlling statistically for the correlation of both with variable  $Z$ .

absence of their individual correlations with  $Z$  (see Table 9.1 for the derivation):

$$r_{(xy).z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}} \quad (9.1)$$

### 9.1.2 Semi-partial correlation

A related expression provides for the *semi-partial correlation* between  $X$  and  $Y$  in the absence of the correlation of  $X$  (the predictor) with  $Z$ :

$$r_{y(x.z)} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1-r_{xz}^2}} \quad (9.2)$$

Whereas the partial correlation coefficient is the simple bivariate correlation of the residuals of both  $X$  and  $Y$  from their predicted values from  $Z$ , the semi-partial correlation is the correlation of  $Y$  with the residuals of  $X$  from its predicted values from  $Z$ .

### 9.1.3 An example of Partial Correlation

As an example<sup>1</sup> of partial correlation, consider the correlation between sales of cigarettes by weight ( $s$ ) and death rate from lung cancer ( $l$ ) over a period of many years ( $t$ )<sup>2</sup>. Because both cigarette sales and the death-rate from lung cancer have increased over time, we would anticipate a correlation between them simply because of this common correlation with time. Suppose the correlation between smoking (i.e., cigarette sales) and year were  $r_{st} = .8$ , that between lung cancer death-rate and year were  $r_{lt} = .875$ , and that the correlation between smoking and lung cancer death rate were  $r_{sl} = .7$ . Plugging these values into the equation for partial correlation (equation 9.1), we find that the partial correlation between smoking and lung cancer death-rate taking the common correlation with time into account is equal to:

$$\begin{aligned} r_{(sl).t} &= \frac{.7 - (.8)(.875)}{\sqrt{1 - .8^2}\sqrt{1 - .875^2}} \\ &= 0.00 \end{aligned}$$

For these (hypothetical, but not unrealistic) values, then, there would be no correlation between smoking and lung cancer death-rate once the common correlation with time is taken into account. But increases in the death-rate from lung cancer are unlikely to occur in the same year as the increases in smoking as lung cancer takes many years to develop, and many more years to reach the point of death. We would expect the increases in death-rate from lung cancer to follow by some years the increases in smoking. Suppose, then, that we *lag* the death-rate by, say, 20 years relative to the smoking data—that is, compute the partial correlation between smoking and the corresponding death-rate from lung cancer 20 years later, again taking the common correlation with time into account. Suppose in doing so that the simple correlations between smoking and lagged-time and lung cancer death-rate and lagged-time not unreasonably remained relatively unchanged at  $r_{st} = .80$  and  $r_{lt} = .85$ , respectively, and that the correlation between smoking and the lung cancer death-rate rose to  $r_{sl} = .9$ . The partial correlation between smoking and lung cancer death-rate lagged by 20 years would be:

$$\begin{aligned} r_{(sl).t} &= \frac{.9 - (.8)(.85)}{\sqrt{1 - .8^2}\sqrt{1 - .85^2}} \\ &= 0.70 \end{aligned}$$

indicating a substantial correlation between smoking and the death-rate from lung cancer 20 years later, even after taking the common correlation with time into account.

<sup>1</sup>To see a provocative discussion using actual data from the years 1888 to 1983 see Peace, L. R. (1985). A time correlation between cigarette smoking and lung cancer. *The Statistician*, 34, 371-381.

<sup>2</sup>We're using  $t$  (for time) rather than  $y$  (for year) to avoid confusing the variables in this example ( $s, l, t$ ) with the variables in the general form of the the partial and multiple correlation formulae ( $x, y, z$ ).

	smoking (s)	year (t)	lung-cancer (l)
smoking (s)	1.0	0.8	0.9
year (t)	0.8	1.0	0.85
lung-cancer (l)	0.9	0.85	1.0

Table 9.2: The correlation matrix of the example variables of smoking (cigarette sales), year, and cancer (lung-cancer death rate 20 years later).

## 9.2 Multiple Correlation

We could ask a different question about three intercorrelated variables. Given that variable  $X$  is correlated individually with both  $Y$  and  $Z$ , what is its *total* correlation with both variables? How much of the variability of  $X$  is “accounted for” by the correlations of  $X$  with both  $Y$  and  $Z$ ? It can’t be just the sum of the individual (squared) correlations with  $Y$  and  $Z$ , because some of the correlation of  $X$  with  $Y$  may be redundant (i.e., correlated with) the correlation of  $X$  with  $Z$ . That is, some of the correlation between  $Y$  and  $Z$  may come about because of their common correlation with  $X$ , so in adding the individual contributions we would be adding this common correlation more than once. The answer to the question is called *multiple correlation*, and is a direct extension of the concept of semi-partial correlation (see section 9.1.2).

### 9.2.1 The Correlation Matrix

We’ll approach the concept of multiple correlation by introducing some new tools. The first of these is the *correlation matrix*—a summary of the intercorrelations between variables in a convenient table format. Table 9.2 summarizes the intercorrelations among the variables in the smoking example from section 9.1.3. The values of 1.0 along the main diagonal of the matrix reflect the correlation of each variable with itself, and the off-diagonal values reflect the correlation between the two variables defining the intersection; hence, inspection of Table 9.2 reveals that year and lung cancer death-rate were correlated  $r_{lt} = .85$ . Note that the upper triangle of the matrix is a simple reflection around the major diagonal of the lower triangle, and is therefore redundant. As a consequence, correlation matrices are often reported as the lower triangle only.

Clearly, correlation matrices could contain the correlations of any number of variables with each other. It is not uncommon in some areas of psychology (e.g., psychometrics, educational psychology) to see published tables containing the correlations between tens and sometimes hundreds of variables! A technique called “Principal Components Analysis” and a related technique called “Factor Analysis” are often used to simplify or to reduce such large matrices to a much smaller number of common structural components or factors. Further discussion of these techniques is beyond the scope of this chapter.

Looking at any row of the correlation matrix, we could ask the multiple correlation question: what is the total correlation of the variable defining the

row with all the other variables? Looking at the three rows of variables in Table 9.2, it appears that some variables (rows) are better “accounted for” than others. For example, the correlations with the remaining variables are generally higher for the row defined by “cancer” than for the rows defined by either “smoking” or “year”. It is this difference in total correlation that we are attempting to capture with the concept of multiple correlation.

### 9.2.2 Euler diagrams or “Ballantine’s”

The second new tool is the use of Euler diagrams or “Ballantine’s”. Euler diagrams are named after their inventor, the mathematical genius Leonhard Euler (1707-1783). They are used to depict the degree to which one logical set is related to another by the degree to which possibly differently-sized (representing the extent of the set) circles overlap with one another. Thus, the proposition “All M are P” could be represented as two completely concentric circles (indicating the complete equivalence of M and P, e.g., unmarried males and bachelors) or as one larger circle completely encompassing a smaller circle (indicating that M, say, is a subset of P, as in “all dogs are mammals”).

In the specific context of multiple correlation, these diagrams are often depicted as three overlapping circles (representing, e.g., the variability of three variables  $X$ ,  $Y$ , and  $Z$ ). Such diagrams are often referred to as “Ballantine’s”, especially in the United States, because three overlapping circles was the logo of the P. Ballantine & Sons brewing company of Newark, New Jersey (see Figure 9.1).<sup>3</sup>

Figure 9.2 depicts the example smoking and cancer correlations as a set of three overlapping circles similar to the Ballantine’s logo. Each circle represents the total variance for a given variable. Because we are working with correlations (which are based on standardised data), the total variance for each variable is 1.0, and the three circles are therefore the same size. The overlap of each circle with another depicts (roughly; our artistry leaves something to be desired) the proportion of common variance between the two variables,  $r_{12}^2$ . The squared correlation between smoking and year for our example data is  $r_{st}^2 = .64$ ; hence, they are depicted as circles with a 64% overlap. Similarly, smoking and lung-cancer are depicted as circles with an 81% overlap, because  $r_{sl}^2 = .81$  for our lagged example data. Finally, lung-cancer is depicted as having an overlap of 72.3% with year, because  $r_{lt}^2 = .723$ .

Imagine that we are interested in explaining the variance in the death-rate from lung cancer as a function of the other two variables. We know that, ignoring year, smoking accounts for 81% of the variance in lung-cancer death-rate, and that year accounts for 72.3%, ignoring smoking. But how much do they *together* account for? Clearly, it is not the simple sum, as that would account for over 100% of the variance, even though, as shown in Figure 9.2, there is still a chunk of the variance in lung-cancer death rate for which neither variable accounts.

---

<sup>3</sup>This is just the first connection between the brewing of beer and statistics; see Chapter 14 for another.



Figure 9.1: The label from one of the brands of beer produced by P. Ballantine & Sons depicting the overlapping circle logo.

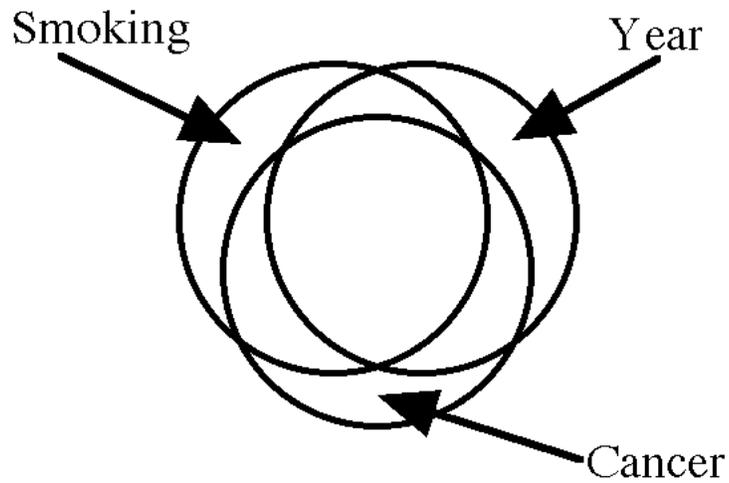


Figure 9.2: A Euler diagram or “Ballantine” of the example correlations from Table 9.2. The area of overlap of any one circle with another represents the squared correlation between the two variables.

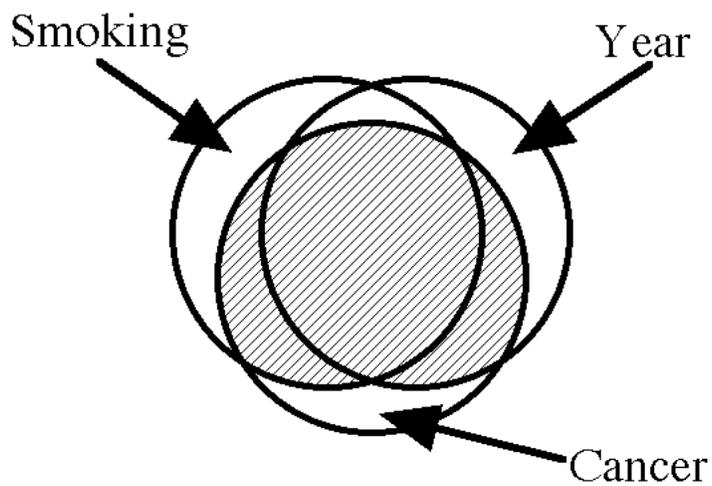


Figure 9.3: The same figure as in Figure 9.2, except shaded to depict the area of interest.

The problem with just adding the two squared, simple correlations together is the area of the lung cancer circle that smoking and year jointly overlap (i.e., the centre polygon in Figure 9.2). The area we want to compute is shown in Figure 9.3.

### 9.2.3 Computing Multiple-R

We can get around the joint overlap problem by starting with either variable, say smoking, and asking what does the remaining variable, year, account for in lung-cancer death rate *beyond* that already accounted for by smoking? That latter proportion is given by the correlation between lung-cancer death-rate and the variable year from which the correlation with smoking has been removed; that is, if we correlated lung-cancer death-rate with the residuals of year from which the predictions from smoking have been removed, we would have the proportion of variance in lung-cancer death rate that is accounted for by year *beyond* (or *independent of*) that accounted for by smoking. This latter value is simply the squared *semipartial* correlation between lung-cancer death rate and year from which smoking has been partialled,  $r_{l(t.s)}^2$  (see section 9.1.2). For the example, using equation 9.2, this value works out to  $r_{l(t.s)}^2 = 0.047$ , or another 4.7% beyond that already accounted for by smoking. Thus, in total, the 81% accounted for by smoking plus the additional 4.7% accounted for by year yields 85.7% of the variance in lung-cancer death-rate accounted for by the combined “effects” of smoking and year. The same final total would have been achieved had we started with the proportion accounted for by year (72.3%), and computed the additional contribution of smoking, (i.e., the squared, semi-partial correlation,  $r_{l(s.t)}^2 = 0.134$ ), for 72.3% plus 13.4% = 85.7%. That is, the squared, *multiple correlation* coefficient—denoted by an uppercase “R”—between lung-cancer death rate, nominally the *dependent variable*, and smoking and year, the *independent or predictor variables*, is  $R_{lst}^2 = 0.857$ .

Unlike the simple, bivariate correlation coefficient, which can be either positive or negative, the multiple correlation coefficient is always positive, varying between 0.0 and 1.0. In the present case, it would be the positive square-root of  $R_{lst}^2 = \sqrt{0.857} = 0.93$ .

Clearly, the process could be continued for any number of variables. For example, imagine a third predictor variable, call it  $F$ .  $R_{lstf}^2$  could be determined by computing, first, the residuals in predicting  $F$  from smoking; call these  $F.S$ . Next, compute the secondary residuals from predicting  $F.S$  from year; call these  $F.ST$ . Last, compute the squared correlation between lung-cancer death-rate and  $F.ST$ , and add this proportion to the earlier  $R_{lst}^2$  to obtain  $R_{lstf}^2$ . Thus, the process can be seen as the sum of successively higher-order squared semi-partial correlation coefficients:  $R_{lstf}^2 = r_{ls}^2 + r_{l(t.s)}^2 + r_{l(f.ts)}^2$ .

There are more mathematically efficient methods available to calculate  $R^2$  between a given dependent variable and any number of independent variables; these are usually presented in the guise of matrix algebra, an approach that is beyond the scope of the level of presentation intended here. However, with  $y$

denoting the dependent variable, and  $1, 2, 3, \dots, K$  denoting the  $K$  predictor or independent variables, the underlying logic of these matrix-algebraic approaches is the same as that just described:  $R_{y123\dots k}^2$  is just the sum of successively higher-order squared semi-partial correlation coefficients.

### 9.3 Questions

- Given the following data of 10 cases measured on 4 variables:

Sex	Age (years)	Like Beer?	Like Scotch?
m	19	y	n
f	26	n	n
f	18	y	n
m	27	n	y
m	20	y	y
m	18	y	n
f	18	y	y
m	25	n	y
m	22	y	y
f	26	n	n

- Construct the correlation matrix of all four variables.
  - What is the *partial correlation* between sex and liking beer after liking scotch is taken into account?
  - What is the *semi-partial correlation* between sex and liking beer after taking age into account for liking beer?
- In your job as a W.H.O. worker you've collected quite a bit of data related to water quality and health in various cities around the world. Your measures of water quality and health are correlated quite highly—Pearson  $r = .75$ . However, being statistically and methodologically astute, you realize that wealth (as measured by average income) may be related to both variables so you correlate income with health and get a Pearson correlation of  $r = .5$ , and you correlate income with water quality and get a Pearson correlation of  $r = .60$ .
    - Is there any relation between water quality and health after you have accounted for the effects of income on water quality?
    - What is the *multiple* correlation,  $R$ , relating health to the independent variables of water quality and income?
  - For a class project, you collect data from 6 students concerning their sex (s), scoring females as 0 and males as 1, grade (g) on a recent statistics (sic) test, and number of beers (b) consumed by him or her the night before the test. You compute the following correlations:  $r_{gs} = -.8325$ ,  $r_{bs} = .665$ ,

and  $r_{gb} = -.83$ . What can you say about the relationship between grade and beer consumption taking sex into account for both variables?

Part II

**Significance**



## Chapter 10

# Introduction to Significance

The origin (and fundamental meaning) of statistical significance and experimental design is usually traced to an apocryphal story told by Sir Ronald Aylmer Fisher in his now-classic, 1935 book, *The design of experiments* (see Figure 10.1).<sup>1</sup> As he tells it, one late, sunny, summer afternoon in Cambridge, England in the late 1920s, a small group of men and women were seated outside at a table for afternoon tea. One of the women insisted that the tea tastes differently depending upon whether the milk or the tea is poured into the cup first.

As the story goes, Fisher, one the the attendees at the tea party, proposed that they test the woman's hypothesis. He suggested that they prepare cups of tea outside of the woman's sight, one-half with milk first, the remainder with tea first, and then present the cups to the woman, one at a time, in a *random* order, recording each of the woman's guesses as to cup-type without feedback. This random assignment of cup-type to ordinal position accomplished two things. First, it assured that, unless the woman actually *had* the ability she claimed to have, her sequence of guesses should not match the actual sequence of cup-types in whole or in part except by chance alone. Second, the random assignment provided a basis for computing these chances. Fisher argued that if the chances that the woman's sequence of guesses matched in whole or in part the actual sequence were sufficiently low, the result should be declared *statistically significant*, and the experiment to have provided support for the woman's hypothesis. His book doesn't record what the actual results of the experiment were.

This idea of *statistical significance* is simple enough. A statement of statistical significance for some aspect of the data is simply a *claim* for the presence of structure supported by some statistical evidence derived from the data. However, the various techniques developed to provide support for such claims from the

---

<sup>1</sup>The story may not be apocryphal. Salsburg (2001) reports that he heard the story from H. Fairfield Smith in the late 1960s, who claimed to have been there. According to Salsburg (2001), Smith also claimed that the woman was perfectly accurate in her guesses.



Figure 10.1: Sir Ronald Aylmer Fisher

data are sometimes complicated, are often controversial, and are not necessarily coherent with one another. Much of the reason for the confusion and controversy can be traced to the history of the development of the techniques, and their subsequent evolution. Over time, often very different and sometimes incommensurate goals and beliefs have been attached to particular techniques such that it is now often not entirely clear what the original intent of a particular method was, or even whether it is capable of meeting both its original and current goals.

Thus, some theorists (the majority) view significance testing as necessarily tied to probability, random sampling, and estimates of population parameters. A statement of statistical significance according to this view is necessarily a statement about one or more populations from the which the current data are (or are not) seen as a random sample. The whole point or purpose of statistical significance by this view is to make *inferences* about populations based on random samples from them. It is these techniques that make up the bulk of most introductory and more advanced textbooks on the theory and application of statistics, as well as the bulk of the mathematical discipline of statistics *per se*. However, possibly because of their ubiquity, it is also these techniques and their interpretations that are the source of much of the historical and current

controversy in statistics.

Other theorists, though, have recognised that especially in disciplines such as psychology, only rarely can the data-sets (never mind the organisms or events generating them) be viewed as random samples from any population. As a consequence, they have argued for a more local claim of structure, that of *effects* in the current experiment or research setting (see, e.g., Edgington, 1995). The claim is still probabilistic, and is intimately tied to the notions of *independence* and *randomisation* (as opposed to but not exclusive of random sampling), but the extension or generality beyond that of the current experiment is *not* part of the claim of statistical significance. Support for such claims of generality must arise from other aspects of the research design (which could include random sampling) or setting.

Most recently, a very small minority of theorists (e.g., Rouanet, Bernard, & Lecoutre, 1986; Vokey, 1998) has argued that statistical significance can be divorced even from probabilistic considerations, as statements merely of the *atypicality* of the data relative to some specified set of possible, potential or theorized results. The idea here, unlike the emphases of the two previous approaches (although, viewed simply as techniques, they embody this principle as well), is that structure is a *relative* concept. Something isn't either structured or not in some absolute sense (viewpoint 1), or only under specially-constructed circumstances (viewpoint 2), but rather is structured or not only relative to something *else*—in this case, some *set* of other possibilities or alternative outcomes: the claim of structure by this viewpoint is not that the data are *patterned*, but that they are patterned *differently*. With careful definitions of the set for comparison, this viewpoint can in fact encompass the previous two as special cases (as we shall see), but neither is seen as *necessary* for a meaningful definition of statistical significance.<sup>2</sup>

The emphasis in this book, then, is the third approach, although it incorporates the *methods* or *techniques* of each of the others. It does so by considering all of the methods as merely means of generating the comparison sets through what are known as *null* or *null hypothesis models* as the *sole* source of *structure* in the data at hand. The question, then, becomes one of whether or not any apparent patterning in the data (e.g., such as the scores in one group being higher on average than those in another group) could reasonably be attributed to the null model—that is, whether the current data could reasonably be seen as typical of the comparison set. If so, the apparent patterning is declared to be *nonsignificant* with respect to that model and comparison set. If not, then the

---

<sup>2</sup>This approach also neatly avoids the very thorny (and as yet unresolved) issue of the definition of a truly random sequence or set, simply by denying the direct need for one. For example, some people would argue that the decimal expansion of the constant  $\pi$  is a random sequence because there is no apparent (to this point in its expansion to millions of digits) patterning of the digits. But others would argue that it is anything but random because simple algorithms can be used to generate every one of the infinite number of digits (hence, the term “pseudo-random” to describe such sequences, including those produced by the “random”-number generators provided with most computer programming languages, which, after all, are just simple, iterative algorithms). This third viewpoint could incorporate such a definition if one ever arises as simply another set of outcomes for comparison.

patterning is declared to be *significant*, in that it is atypical of what one would expect according to the null model. That is, to declare some pattern *significant* is to *reject* a given null model as a source of the patterning.

Fisher intended the term “significant” with the everyday meaning it had at the turn of the 20th century, namely, “salient”, “stands out”, “decidedly different”, or “markedly deviant”. Since then, the everyday meaning of “significant” has shifted so that the principal meaning is now one of “important”, leading to the often unfortunate interpretation that a statistically significant result is, thereby, important. Not necessarily. Sometimes a statistically significant result is important, such as when it suggests a cure for cancer, other times not, as when it merely confirms a truism. Sometimes, it is the lack of significance that is important, as when a commonly accepted treatment, such as chiropractic or homoeopathy, is found to be of no value.

## 10.1 The Canonical Example

The canonical case for a test of significance is the comparison between two groups of observations. Virtually every statistics textbook focuses the discussion with such an example, and the test for a significant difference between two groups is probably the most commonly used form of statistical significance test. As intimated by the previous discussion of the different viewpoints, there is not just one test for a significant difference between groups, or even just one *kind* of test. Rather, there is a whole *class* of such tests with its members emphasising different aspects of the data, different statistics derived from the data, different assumptions, different purposes, different comparison sets, and so on. Because of all these differences, no one test is universally better than all the others, although for some purposes and some statistics, some tests are clearly better and some clearly worse than others. Before we can explore this canonical case, however, we need some basic tools to assist us in counting events.

### 10.1.1 Counting Events

#### The fundamental counting rule

In a series of  $N$  independent events, if event 1 can occur  $k_1$  ways, event 2 can occur  $k_2$  ways, and so on, then the total number of ways the  $N$  events can occur is:

$$k_1 * k_2 * \dots * k_N$$

So, for example, if there are 2 ways to flip a coin and 52 ways to draw a card from a deck of cards, then the total number ways the coin can be flipped and a card drawn is  $2 * 52 = 104$ .

If the events are all of the same type (e.g., the flipping of multiple coins or, equivalently, flipping the same coin multiple times), the fundamental counting rule reduces to:

$$k^N$$

### Permutations

A common counting problem is to determine the number of different *orders* or *permutations* in which  $N$  different events or objects can be arranged. For example, if you had 6 family members, you might wish to know in how many different orders the family members could be arranged in a line for a family photograph. The answer is that there are 6 ways the first position could be filled, 5 ways (from the remaining 5 family members after the first position is filled) that the second position could be filled, 4 ways that the third position could be filled, and so on. Applying the fundamental counting rule, those values yield a total of  $6 * 5 * 4 * 3 * 2 * 1 = 720$ .

In general,  $N$  events can be permuted in  $N!$  (read “ $N$ -factorial”) ways.<sup>3</sup> That is, the number of ways  $N$  events can be permuted taken in groups of  $N$  at a time is:

$$P_N^N = N!$$

where

$$N! = N * (N - 1) * \cdots * (N - (N - 1))$$

$N$ -factorial is perhaps more easily defined recursively. Defining  $0! = 1$ , then

$$N! = N * (N - 1)!$$

Thus, the factorial of 6,  $6!$ , for example, could be written as:

$$6! = 6 * 5! = 6 * 5 * 4! = \cdots = 6 * 5 * 4 * 3 * 2 * 1 * 0! = 720$$

Suppose you were interested, instead, in how many different photographs could be taken of the 6 family members if only three were to be included each time. According to the fundamental counting rule, there are 6 ways the first position could be filled, 5 ways the second could be filled, and 4 ways the third could be filled, for a total of  $6 * 5 * 4 = 120$  different photographs, which is just the original  $6!$  divided by the  $3!$  of the original last 3 positions no longer being filled. Thus, in general, the number of ways of permuting  $N$  things taken  $r$  ( $r \leq N$ ) at a time is:

$$P_r^N = \frac{N!}{(N - r)!}$$

Note that  $P_N^N = P_{(N-1)}^N$ .

### Combinations

Suppose you were interested in how many different photographs could be taken of the 6 family members if only 3 were to be included each time *and* the order of the three within the photograph was of no concern. That is, how many different *sets* or *combinations* of 3 family members could be made from the 6 family

---

<sup>3</sup>It is fun to think that the symbol for factorial numbers, the exclamation mark, was chosen to reflect the fact that these numbers get large *surprisingly* quickly — a mathematical witticism.

members? Note that for every set of 3 family members, there are 3! ways of permuting them. We are concerned only with the number of *different* sets, not the ordering within the sets. Thus, as  $P_r^N$  yields the total number of different permutations, and  $r!$  is the number of permutations per *set* of  $r$  things, then the number of different *sets* or *combinations* is given by:

$$C_r^N = \frac{P_r^N}{r!}$$

For our example case, then, the number of different sets of 3 family members that can be made from 6 family members is the number of permutations,  $6!/3! = 120$ , divided by the number of permutations of any one of those sets,  $3! = 6$ , yielding  $120/6 = 20$  different sets.

Mathematicians commonly use a different symbol for combinations. Coupling that symbol with the expression for  $P_r^N$ , yields the following general equation for the number of combinations of  $N$  things taken  $r$  at a time (read as “N choose r”):

$$\binom{N}{r} = \frac{N!}{r!(N-r)!}$$

### 10.1.2 The Example

Suitably armed with some basic counting tools, we can now approach the canonical example of a test of significance of the comparison between two groups of observations. Consider, as a focal example, two groups of 4 scores each from, say, 4 males and 4 females, although *any distinction* (e.g., capricorns and leos, Montagues and Capulets, blondes and brunettes, tea-first and milk-first, controls and treated) will do. For convenience, we’ll assume the scores are all numerically different.

Before knowing anything else about the 8 scores, if we simply rank-order the 8 scores, how many different orders considering only the group labels are possible? That is, how many different patterns of the group labels can result from rank-ordering the corresponding scores? One order, for example, could be MMFMFFFM, another could be MFMFFMFM, and so on. As there are 8! different ways of ranking the scores, but 4! of those are simply rearrangements of the males, and another 4! of them are simply rearrangements of the females, then the total number of unique patterns is given by:

$$\frac{8!}{4!4!} = 70$$

The question we wish to ask is whether the pattern we actually obtained is consistent with the idea that the male scores are *exchangeable* with the female scores. That is, whether the pattern we obtained is *typical* of the kinds of patterns one observes by simply rearranging the group labels. Most patterns will be similar to the two suggested earlier: MMFMFFFM and MFMFFMFM, in which the males and females appear thoroughly intermixed, and, hence, interchangeable with one another. Suppose though that the pattern of group labels we actually

obtained when rank-ordered by the corresponding scores was FFFFMMMM; that is, that the top four scores all belonged to the males. Is it reasonable to consider this pattern to be *typical* of the 70 different patterns possible? Of the 70, the pattern FFFFMMMM only occurs once; but then, that is also true of any of the other patterns. But, patterns *like* it also only occur once: there is only one way to have all the males score higher than all the females, whereas, for example, there are an additional 16 different ways to have 3 of the male scores among the top 4. (There are 4 ways the low male could be among the 3 low females, and 4 ways the high female could be among the 3 high males, for a total of  $4 \times 4 = 16$  different patterns.) Thus, relative to many of the other patterns, FFFFMMMM is unusual or *atypical*. It stands out; it is *significant*. It suggests that the male scores are *not* interchangeable with the female scores.

How atypical is atypical enough? Conventional wisdom<sup>4</sup> has it that if fewer than 5% (1 in 20) of the possible patterns or potential cases can be considered to be like the one in question, then the result is considered atypical of the reference set, and declared *statistically significant*.<sup>5</sup> Thus, the top 4 scores being all male is considered significant ( $1/70 < 1/20$ ), but 3 or more of the top scores being male is not (all male can happen only one way, and 3 males can happen 16 ways, for a total of 17 ways:  $17/70 > 1/20$ ). Note that these fractions are not (necessarily) probabilities; we refer to them (as did Fisher) as *p-values* to be neutral on the issue of whether they can properly be considered probabilities, or simply proportions of some designated set. The distinction between the two is easily seen: if the scores in our example were *heights* or *weights*, the result that the top four scores were all males is still atypical of the set of patterns even though the result is highly probable (for random samples of males and females).

### 10.1.3 Permutation/Exchangeability Tests

Suppose we wished to use more of the information inherent in the scores than just their ordinal values. For example, it wasn't just that the male scores were ordinally higher than the female scores, they were numerically higher—where *numerically* is the term we wish to cash in. To make the example concrete, suppose that the data were that the males had scores  $\{4, 6, 7, 8\}$ , and the females had the scores  $\{1, 2, 3, 5\}$ : that is, only 3 of the 4 male scores were among the top 4—and so would not be considered an atypical pattern (and not thereby significant) by the approach just discussed, but in general or *on average* the male scores exceeded the female scores.

If the male and female scores were truly exchangeable, then it shouldn't matter (much) if we in fact exchange the scores across the male/female distinction. The mean difference (to use just one of many statistics we could have chosen) we

<sup>4</sup>A lovely phrase invented by John Kenneth Galbraith that has quickly entered the language—evidence that the language clearly needed it. It is meant to be used disparagingly.

<sup>5</sup>Fisher never specified a specific criterion for statistical significance, apparently believing that the precise value should be left up to the judgement of the individual scientist for any given comparison. Some subsequent theorists believe this fluidity to be anathema, and demand that the criterion for significance not only be specified, but specified in advance of collecting the data!

Males	Females	Male Mean	Female Mean	Mean Difference
4,6,7,8	1,2,3,5	6.25	2.75	3.5
5,6,7,8	1,2,3,4	6.5	2.5	4
4,5,7,8	1,2,3,6	6	3	3
4,5,7,3	1,2,8,6	4.75	4.25	0.5
1,2,8,7	4,6,5,3	4.5	4.5	0
4,3,2,1	8,7,6,5	2.5	6.5	-4
5,3,2,1	8,7,6,4	2.75	6.25	-3.5

Table 10.1: A sampling of the 70 possible exchanges of the male and female scores.

actually obtained between the two groups (i.e., 6.25 for the males minus 2.75 for the females = 3.5) should not be all that different from what we would generally obtain by exchanging scores between the sexes. As before, there are 70 different ways<sup>6</sup> we can exchange the scores between the two groups.<sup>7</sup> A few examples of these exchanges, and the resulting means and mean differences are shown in Table 10.1.

A histogram of the distribution of the mean differences from the complete set of 70 exchanges is shown in Figure 10.2. Of these, only the original distribution of scores and one other (the one that exchanges the high female for the low male) produces a mean difference as large or larger than the obtained difference—a proportion of only  $2/70 = 0.028571$ . Hence, we would declare the result *significant*; the mean difference obtained from this distribution is *atypical* of those that result from exchanging male and female scores. Said differently, the male and female scores appear *not* to be exchangeable, at least with respect to means.

The set or *population* of patterns to which we wish to compare our particular result can be generated in innumerable ways, depending on how we wish to characterise the process with respect to which we wish to declare our result typical or atypical. In the next chapter, we consider a theoretical distribution of patterns of binary events that can be adapted to many real-world situations.

## 10.2 Questions

1. From 8 positive and 6 negative numbers, 4 numbers are chosen without replacement and *multiplied*. What proportion of all sets of 4 numbers

<sup>6</sup>This number represents the number of unique exchanges—including none at all—of scores between each of the 4 males and each of the 4 females. Usually, a computer program is used to determine the exchanges; hence, the need to await the easy access to computers for this fundamental method to be a practical approach to statistics.

<sup>7</sup>We could also exchange scores within the groups, but that would not affect the group means, and, hence, would not influence the difference between the means—our statistic of choice here, nor would it change the *proportion* of scores equal to or greater than our obtained difference.

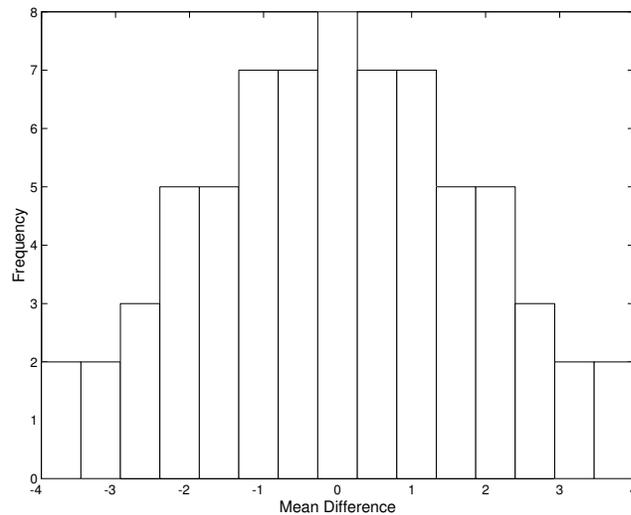


Figure 10.2: The frequency distribution of the 70 mean differences between the male and female sets of scores produced from the possible 70 unique exchanges of male and female scores.

chosen this way will the product be a *positive number*?

2. A die is tossed and a card is chosen at random from a normal (52-card) deck. What proportion of all the ways of doing this
  - (a) will the die show a number *less than 3* OR the card is from a *red suit*?
  - (b) will the die show a number *less than 3* AND the card is from the *diamond suit*?
3. Two dice are rolled. Given that the two faces show different values, what proportion of all such rolls will one of the faces be a 3?
4. In a group containing 10 males and 20 females, 3 people are chosen without replacement. Of all such ways of selecting 3 people from this group, what proportion would contain no more than 1 male?
5. In a group containing 5 males and 8 females, 2 people are chosen without replacement. Of all the pairs possible, what proportion will be a same-sex couple?
6. Ten people in a room are wearing badges marked 1 through 10. Three people are chosen to leave the room simultaneously. Their badge number is noted. What proportion of such 3 people selections would

- (a) have the *smallest* badge number equal to 5?
  - (b) have the *largest* badge number equal to 7?
7. A die is thrown  $n$  times. Of all the ways of throwing a die  $n$  times, what proportion will have *at least* one “6” come up?
8. Each of two people tosses three coins. Of all the results possible, what proportion has the two people obtaining the same number of heads?

## Chapter 11

# The Binomial Distribution

Consider what is known as a binary or binomial situation: an event either is or has property  $X$  or it isn't or doesn't; no other possibilities exist, the two possibilities are *mutually exclusive* and *exhaustive*. Such distinctions are also known as *contradictories*: event 1 is the contradiction of event 2, not just its contrary. For example, blue and not-blue are contradictories, but blue and orange are merely contraries.

In a set of these events, suppose further that  $X$  makes up the proportion or has the relative frequency  $p$  of the set, and not- $X$  makes up the remaining proportion or has relative frequency  $1 - p = q$  of the set. Clearly, then,

$$p + q = 1$$

So, if we selected one of these events at random (i.e., without respect to the distinction in question) then we would expect to see the  $p$ -event  $100p\%$  of the time, and the  $q$ -event  $100q\%$  of the time—that is, this result would be our expectation if we repeated this selection over and over again *with replacement*. With replacement means that we return the event just selected to the pool before we select again, or more generally that having selected an event has no effect on the subsequent likelihood of selecting that event again—the selections are *independent*.

Now, what if we selected 2 of these events at random with replacement, and repeated this selection of two events over many trials? Clearly, there are 3 possibilities for any such selection trial. We could observe 2  $p$ -events, 1  $p$ -event and 1  $q$ -event, or 2  $q$ -events. At what rates would we expect to observe each of these possibilities? If we square  $(p + q)$  to represent the two selections per trial, we get:

$$(p + q)^2 = p^2 + 2pq + q^2 = 1$$

This expression may be rewritten as:

$$(p + q)^2 = (1)p^2q^0 + (2)p^1q^1 + (1)p^0q^2 = 1$$

Note that it conveniently partitions the total probability (of 1.0) into three summed terms, each having the same form. Before discussing this form further, let's see what happens if we *cube*  $(p + q)$ , representing 3 selections per trial:

$$(p + q)^3 = (1)p^3q^0 + (3)p^2q^1 + (3)p^1q^2 + (1)p^0q^3 = 1$$

And so on.

Each term in each of these expansions has the following form:

$$\binom{N}{r} p^r q^{N-r}$$

where  $\binom{N}{r}$  provides the number of different ways that particular combination of  $r$   $p$ -events and  $N - r$   $q$ -events can happen if there are a total of  $N$  such events, and  $p^r q^{N-r}$  provides the *probability* of that particular combination over the  $N$  events. For example, with 3 selections per trial, obtaining 3  $p$ -events (and, hence, 0  $q$ -events) can occur only 1 way, with the probability of that way of  $ppp = p^3q^0$ , whereas 2  $p$ -events and 1  $q$ -event can occur 3 ways (i.e.,  $ppq$ ,  $pqp$ , and  $qpp$ ), with the probability of any one of those ways of  $ppq = p^2q^1$ . Thus, the probability of observing exactly 2  $p$ -events in 3 independent selections is  $3p^2q^1$ .

Consider now a *population* (or relative frequency or probability) distribution,  $X$ , consisting of two values:  $X = 1$  with probability  $p$ , and  $X = 0$  with probability  $q = 1 - p$ . Suppose we construct the *statistic*  $Z = \sum_{i=1}^N X_i$ , sampling independently from  $X$ ,  $N$  times; the statistic  $Z$  is just the frequency of  $X = 1$  for a given trial of  $N$  selections, which can range from 0 to  $N$ . What is the *sampling distribution* of  $Z$ ? That is, how are the different values of  $Z$  distributed? From the foregoing, the relative frequency of any particular value,  $r$ , of  $Z$  ( $0 \leq Z \leq N$ ) is given by:

$$p(Z = r) = \binom{N}{r} p^r q^{N-r} \quad (11.1)$$

An example of a distribution of all 11 ( $0 \leq Z \leq 10$ ) such values of  $Z$  for  $N = 10$  and  $p = .5$  is shown in Table 11.1 and Figure 11.1. Such distributions, based on equation 11.1 are called *binomial distributions*. A binomial distribution is a model of what is known as a *Bernoulli process*: two mutually exclusive and exhaustive events with independent trials.

Consider the following simple application of the binomial distribution. Normal coins when tossed, are often supposed to be *fair*—meaning that the coin (via the tossing procedure) is just as likely to come up heads as tails on any given toss, and coming heads or tails on any given toss in no way influences whether or not the coin comes up heads or tails on the next toss.<sup>1</sup> We can use the binomial distribution shown in Table 11.1 as a *model* of such a *fair* coin tossed 10 times. That is, the distribution shown in Table 11.1 and Figure 11.1 provides the relative frequency of observing any given number of heads in 10 tosses of a fair coin.

<sup>1</sup>It is also assumed that the coin can *only* come up heads or tails (e.g., it can't land on edge), and only one of the two possibilities can occur on any given toss, i.e., a *Bernoulli process*.

Value of $Z$										
0	1	2	3	4	5	6	7	8	9	10
.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
.001	.011	.055	.172	.377	.623	.828	.945	.989	.999	1
1	.999	.989	.945	.828	.623	.377	.172	.055	.011	.001

Table 11.1: The binomial distribution ( $X = \{0, 1\}$ ) of  $Z = \sum X$  for  $N = 10$  trials,  $p = .5$ . Each column provides the relative frequency, the cumulative relative frequency (up), and the cumulative relative frequency (down) for the respective values of  $Z$  from 0 to 10. The relative frequencies (i.e., the values in the first row) are plotted in Figure 11.1.

Suppose, then, that we observe 9 heads in 10 tosses of the coin, and we ask: Is the coin fair? According to our *model* of a fair coin, 9 or more<sup>2</sup> heads should occur less than 1.1% of the time (see Table 11.1). Now, we can accept one of two alternatives: Either we have just observed one out of a set of events whose combined relative frequency of occurrence is only 1.1%, *or* there is something wrong with our model of the coin. If 1.1% strikes us as sufficiently improbable or unlikely, then we *reject* the model, and conclude that something other than our model is going on. These “other things” could be any one or more of an infinite set of possibilities. For example, maybe the model is mostly correct, except that the coin (or tossing procedure) in question is *biased* to come up heads much more frequently than tails. Or maybe the trials aren’t really independent: maybe having once come up heads, the coin (or the tossing procedure) just continues to generate heads. And so on.

### 11.0.1 Type I and Type II errors

Of course, we could also just be *wrong* in rejecting the model. The model may in fact be a perfectly valid description of what we observed; we just happened to encounter a rare (to us) event allowed by the model. Still, we would only make this error 1.1% of time. That is, the probability that we have mistakenly rejected the model is only .011.

#### One and two-tailed tests

Actually, to be consistent, we would probably also reject the model if we had observed an unlikely few number of heads (say, 1 or fewer). So, we would probably perform what is known as a *two-tailed test*, rejecting the model if we observed too many or too few heads, for a combined probability of incorrectly rejecting the model of  $.011 + .011 = 0.22$  or 2.2%. If we had a reason for emphasizing only

<sup>2</sup>We include the “or more” because if we were to find 9 heads in 10 tosses of the coin impressive, we should be even more impressed by even more extreme possibilities, such as 10 heads in 10 tosses. Thus, we want the combined relative frequency of the whole set of results that we would find impressive should they occur.

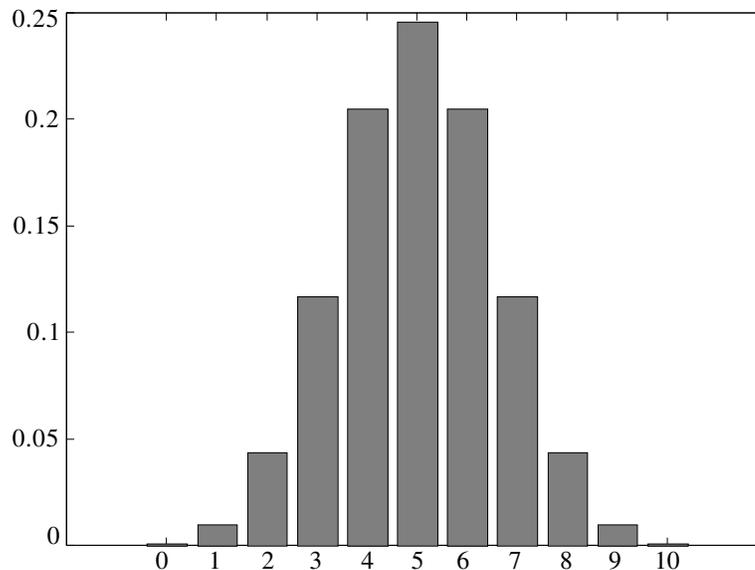


Figure 11.1: Relative frequency as a function of the statistic  $Z$ . A plot of the binomial distribution from Table 11.1.

one or the other tail of the distribution, such as the investigation of a directional prediction (i.e., we were concerned that the coin was biased in favour of heads; hence, either alternative, being unbiased—the model—or biased against heads would suffice to disconfirm the prediction), we would probably prefer to use the *one-tailed test*.

### Type I errors and $\alpha$

Incorrectly rejecting the model is the first of the errors we could make, and for that reason is called a *Type I error*. It is conventional to use the first letter of the Greek alphabet, the lower-case symbol  $\alpha$ , to denote the probability of the error of incorrectly rejecting the model. So, for our current example case, we would say that our “alpha-level” (i.e., probability of incorrectly rejecting the model) was  $\alpha = .022$  because if the model were true that is the proportion of times we would expect to incorrectly reject it with this result. It is also conventional (at least within psychology) to use an alpha-level of  $\alpha = .05$  as the criterion in deciding whether or not to reject the model.

Note, however, that these are *conditional* probabilities. They reflect the probability of rejecting the model incorrectly *given that the model is the correct description* of the process that gave rise to the data. What we don’t know, and what significance testing *cannot* tell us is the probability of the conditional itself—the probability that the model is in fact true. All we get (and, in fact,

we determine because we decide whether or not to reject the model at a specific alpha-level) is the probability that we will mistakenly reject the model assuming it is true.

### Type II errors and $\beta$

A related error is *failing to reject the model when we should*; that is, failing to reject the model when it is not, in fact, a valid description of the process that gave rise to the data. As this is the second of the errors we can make, it is referred to as a *Type II error*, and the symbol for the probability of making a Type II error is the second letter of the Greek alphabet,  $\beta$ . We can know even less about this probability than we can about alpha because there are myriad, often infinite, ways in which the model could be invalid, ranging from it being *almost* correct (e.g., the model is accurate except that  $p$  actually equals .6 not .5), to it being wrong in every way one could imagine.

## 11.0.2 Null and Alternative Hypotheses, and Power

By making the assumption (or argument) that the model (such as the binomial) *is* valid except for possibly the specification of one of the model's parameters, the general significance test can be turned into a test of the value of that specific parameter. Application of the model in this case amounts to a test of the hypothesis that the parameter in question has or doesn't have the specified value. For example, in the coin-tossing situation we could argue that we have every reason to believe that the coin tosses are indeed independent, Bernoulli trials as required by the binomial distribution model, but that we suspect that the coin is biased somehow to come up, say, heads more often than tails. If we then compare our obtained results to the binomial model with  $p = .5$  we would be testing the hypothesis that in the hypothetical population of coin tosses from which ours were but a sample, the value of the parameter corresponding to  $p$  is equal (or not) to .5.

The bulk of the statistical tests described in most statistics texts are of this type, and are often described and developed explicitly as tests of specific parameters of populations.<sup>3</sup> Under these circumstances, given that the assumptions are valid,  $\beta$ , the probability of failing to reject the model when we should, can be defined. The model from which the requisite probability can be calculated is simply the original model—referred to as the “null model” or “null hypothesis”—with the parameter changed to some alternative value. This altered model is referred to as the “alternative hypothesis”. The probability, then, of the contradictory event,  $1 - \beta$ , is referred to as the *power* of the statistical test for that parameter, and is a measure of the *sensitivity* of the test to detect such specific differences between the null and alternative models. Again, however, even where calculable,

---

<sup>3</sup>It should be clear that population parameter testing is *not* a necessary characteristic of significance tests. Indeed, in the absence of random sampling from said populations, significance tests in general could not meaningfully be so characterized.

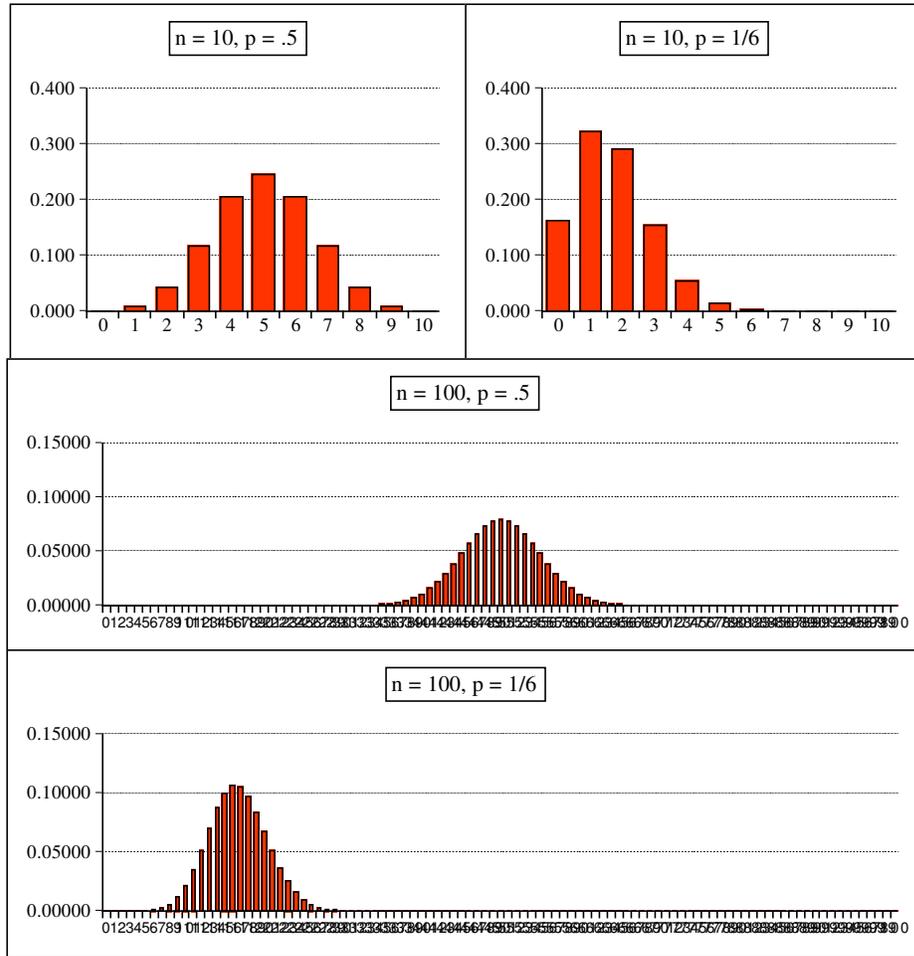


Figure 11.2: Examples of binomial distributions. Note that the distributions become more symmetrical either as  $p$  approaches  $.5$ , or  $n$  increases.

these are not *absolute* probabilities: they are still conditional on the assumptions' being a valid description of the process that generated the data.

In psychology and most other biological and life sciences it is rare that our theories are so well-specified that the statistical models and values of the parameters can be directly derived from them. Typically, we get around this problem by either (1) adapting our data collection process to match as closely as possible to that of the null model (e.g., random assignment), or (2) using models and techniques that are as general as possible (e.g., random sampling and the testing of parameters that are likely to be distributed in predictable ways), or (3) typically both. These approaches are the subject of the next chapter.

### 11.0.3 The Sign Test

Sometimes, though, we can simply *reconceptualise* a particular statistical question about a given data-set so that it can be seen as a version of a pre-existing model. One classic example of this approach to significance testing is known as the *sign test*, an application of models based on the binomial distribution. For example, suppose for a given set of data that you were interested in whether the scores *improved* over two attempts, such as a midterm exam and a final. We could use the permutation techniques discussed in Chapter 10 and compare the obtained mean difference, say, to the complete distribution of such mean differences obtained from all combinations of swapping each midterm score with its paired final score.

But the binomial distribution suggests another approach. *If* there was no general trend toward improvement, then improvement should happen about as often as no improvement. That is, as the null model, we could view each case as a Bernoulli trial with the probability of improvement,  $p = .5$ , the same as no improvement. We could then simply count the number of cases that actually did improve,  $z$ , out of the total number of cases,  $n$ , and compute the probability of obtaining  $z$  or greater if the binomial model with  $n$  trials and  $p = .5$  were true, rejecting the model if the probability were sufficiently low. In fact, any situation that could be reconceptualised as having independent, positive or negative trials (hence the name “sign-test”) could be modelled in this way. The “Pepsi Challenge” is one obvious example in which the sign-test would be an appropriate null model (and, incidentally, is rejected about as often as the model with  $\alpha = .05$  suggests it should be: about 5% of the time—which suggests, of course, that people really can't discriminate Coke from Pepsi in blind taste tests).

## 11.1 Questions

1. Listed below are the distances in centimetres from the bulls-eye for inebriated bar patrons' first and second attempts to hit the centre of the dart board. Use the sign-test to test (at the  $\alpha = .05$  level) the hypothesis that accuracy improved over the two attempts:

patron	Throws	
	First	Second
a	5	1
b	8	9
c	2.5	1.3
d	6.8	4.2
e	5.2	2
f	6.7	4
g	2.2	1.1
h	.25	.1
i	4.3	4.2
j	3	2

2. A coin is tossed 12 times, and 10 heads are observed. Based on the evidence, is it likely to be a fair coin (i.e., where “fair coin” includes the tossing process)?
3. At the University of Deathbridge, the students’ residence is infested with *Bugus Vulgaris*. These nasty insects come in two sizes: large and extra-large. In recent years, the extra-large appear to have gained the upper hand, and out-number the large by a ratio of 3:1. Eight *Bugus Vulgaris* have just walked across some poor student’s residence floor (carrying off most of the food and at least one of the roommates, in the process). Assuming (not unreasonably) that the number of *Bugus Vulgaris* in the residences borders on infinite, what is the probability that *no more* than two of the eight currently crossing the residence floor are extra-large?
4. The probability that Mike will sleep in and be late for class is known to be (from previous years and courses)  $p = .5$ . Lately, though, you’ve begun to wonder whether Mike’s being late for this particular class involves more than simply the random variation expected. You note that the number of times Mike was late over the last 10 classes was 9. Assuming that Mike either is or isn’t late for class and that his attempts to get to class on time are independent, do you have reason to suspect that something is going on beyond the normal random variation expected of Mike? Explain why or why not.
5. The probability of getting an “B” or better in sadistics (sic) is only .40 (that’s why it’s called *sadistics*). Eight sadistics students are selected at random. What is the probability that *more than*  $2/3$  of them fail to get a “B” or better in sadistics?
6. At the U of L, 20% of the students have taken one or more courses in statistics.
  - (a) If you randomly sample 10 U of L students, what is the probability that your sample will contain more than 2 students who have taken statistics?

- (b) If you randomly sample 7 U of L students, what is the probability that none will have taken statistics OR all will have?



## Chapter 12

# The Central Limit Theorem

As we've seen, the binomial distribution is an easy-to-use theoretical distribution to which many real-world questions may be adapted. But as the number of trials or cases gets larger, the computations associated with calculating the requisite tail probabilities become increasingly burdensome. This problem was recognised as early as 1733 by Abraham De Moivre who asked what happens to the binomial distribution as  $n$  gets larger and larger? He noted that as  $n$  increases the binomial distribution gets increasingly smooth, symmetric (even for values of  $p \neq .5$ ), and “bell-shaped”. He found that “in the limit”, as  $n$  goes to infinity, the binomial distribution converges on a continuous distribution now referred to as the “normal distribution”.<sup>1</sup> That is, as you sum over more and more independent Bernoulli events, the resulting distribution is more and more closely approximated by the normal distribution.

### 12.1 The Normal Distribution

The relative frequency distribution (or, as mathematicians are wont to say, the probability density function, *pdf*) of the normal distribution is given by the following rather complex-looking expression:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (12.1)$$

in which  $\mu$  and  $\sigma$  are the mean and standard deviation, respectively. If we *standardise* the distribution on  $\mu$  and  $\sigma$ , we get the *pdf* of the *standard normal*

---

<sup>1</sup>It is also referred to, although mostly only by mathematicians, as the “Gaussian distribution”, after Carl Friedrich Gauss who along with Laplace and Legendre in the early 1800s recognised its more general nature. In fact, the naming of this distribution is a prime example of *Stigler's Law of Eponymy* (Stigler, 1999), which states that no discovery in math (or science) is named after its true discover (a principle that even applies, recursively, to Stigler's Law!).

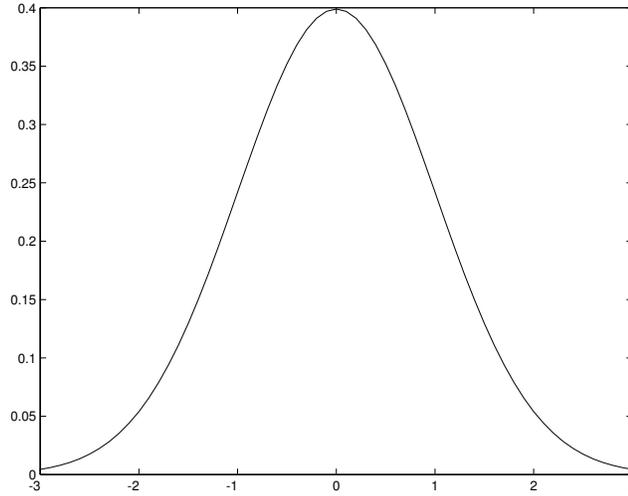


Figure 12.1: The standard normal curve with  $\mu = 0$  and  $\sigma = 1$ .  $\phi(x)$  is plotted as a function of standard deviation units from the mean.

*distribution*, usually symbolized as  $\phi(x)$ :

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (12.2)$$

A plot of this standard normal distribution is shown in Figure 12.1.

To be useful as an approximation to the binomial distribution, we need to extract probabilities from this function. As it is a continuous distribution, these probabilities are defined as the proportion of area under the curve for various ranges, and are computed by *integrating* (a variant of summation for continuous distributions) between two different values along the x-axis using the integral calculus. These areas or proportions of the total curve are usually denoted as  $\Phi(X)$ . Table 12.1 summarises these areas or values of  $\Phi(X)$  for a series of values of standard scores, in increments of 0.01, ranging from zero to 4 standard deviations from the mean. For each standard or  $z$ -score, two proportions are given: the proportion of the total area between the mean and the  $z$ -score, and the area beyond the  $z$ -score. So, for example, 34.13% of the scores from a normal distribution are between the mean and 1 standard deviation above the mean, and 15.87% of the scores are beyond 1 standard deviation above the mean. As the distribution is symmetrical about the mean, the same proportions are found for 1 standard deviation below the mean. Thus, 68.26% of the scores from a normal distribution are within 1 standard deviation of the mean, and 31.74% of them are beyond it. Similarly, a  $z$ -score of 1.65 is needed to leave a tail with an area of 5%, and a  $z$ -score of  $\mp 1.96$  (roughly 2 standard deviations) is needed to leave 2.5% in both tails, for a total of 5%.

Table 12.1: Areas  $[\Phi(z)]$  under the Normal Curve

$z$	$\mu$ to $z$	beyond $z$	$z$	$\mu$ to $z$	beyond $z$
0	0	.5	.4	.1554	.3446
.01	.004	.496	.41	.1591	.3409
.02	.008	.492	.42	.1628	.3372
.03	.012	.488	.43	.1664	.3336
.04	.016	.484	.44	.17	.33
.05	.0199	.4801	.45	.1736	.3264
.06	.0239	.4761	.46	.1772	.3228
.07	.0279	.4721	.47	.1808	.3192
.08	.0319	.4681	.48	.1844	.3156
.09	.0359	.4641	.49	.1879	.3121
.1	.0398	.4602	.5	.1915	.3085
.11	.0438	.4562	.51	.195	.305
.12	.0478	.4522	.52	.1985	.3015
.13	.0517	.4483	.53	.2019	.2981
.14	.0557	.4443	.54	.2054	.2946
.15	.0596	.4404	.55	.2088	.2912
.16	.0636	.4364	.56	.2123	.2877
.17	.0675	.4325	.57	.2157	.2843
.18	.0714	.4286	.58	.219	.281
.19	.0753	.4247	.59	.2224	.2776
.2	.0793	.4207	.6	.2257	.2743
.21	.0832	.4168	.61	.2291	.2709
.22	.0871	.4129	.62	.2324	.2676
.23	.091	.409	.63	.2357	.2643
.24	.0948	.4052	.64	.2389	.2611
.25	.0987	.4013	.65	.2422	.2578
.26	.1026	.3974	.66	.2454	.2546
.27	.1064	.3936	.67	.2486	.2514
.28	.1103	.3897	.68	.2517	.2483
.29	.1141	.3859	.69	.2549	.2451
.3	.1179	.3821	.7	.258	.242
.31	.1217	.3783	.71	.2611	.2389
.32	.1255	.3745	.72	.2642	.2358
.33	.1293	.3707	.73	.2673	.2327
.34	.1331	.3669	.74	.2704	.2296
.35	.1368	.3632	.75	.2734	.2266
.36	.1406	.3594	.76	.2764	.2236
.37	.1443	.3557	.77	.2794	.2206
.38	.148	.352	.78	.2823	.2177
.39	.1517	.3483	.79	.2852	.2148

*continued on next page*

Table 12.1 continued from previous page

$z$	$\mu$ to $z$	beyond $z$	$z$	$\mu$ to $z$	beyond $z$
.8	.2881	.2119	1.2	.3849	.1151
.81	.291	.209	1.21	.3869	.1131
.82	.2939	.2061	1.22	.3888	.1112
.83	.2967	.2033	1.23	.3907	.1093
.84	.2995	.2005	1.24	.3925	.1075
.85	.3023	.1977	1.25	.3944	.1056
.86	.3051	.1949	1.26	.3962	.1038
.87	.3078	.1922	1.27	.398	.102
.88	.3106	.1894	1.28	.3997	.1003
.89	.3133	.1867	1.29	.4015	.0985
.9	.3159	.1841	1.3	.4032	.0968
.91	.3186	.1814	1.31	.4049	.0951
.92	.3212	.1788	1.32	.4066	.0934
.93	.3238	.1762	1.33	.4082	.0918
.94	.3264	.1736	1.34	.4099	.0901
.95	.3289	.1711	1.35	.4115	.0885
.96	.3315	.1685	1.36	.4131	.0869
.97	.334	.166	1.37	.4147	.0853
.98	.3365	.1635	1.38	.4162	.0838
.99	.3389	.1611	1.39	.4177	.0823
1	.3413	.1587	1.4	.4192	.0808
1.01	.3438	.1562	1.41	.4207	.0793
1.02	.3461	.1539	1.42	.4222	.0778
1.03	.3485	.1515	1.43	.4236	.0764
1.04	.3508	.1492	1.44	.4251	.0749
1.05	.3531	.1469	1.45	.4265	.0735
1.06	.3554	.1446	1.46	.4279	.0721
1.07	.3577	.1423	1.47	.4292	.0708
1.08	.3599	.1401	1.48	.4306	.0694
1.09	.3621	.1379	1.49	.4319	.0681
1.1	.3643	.1357	1.5	.4332	.0668
1.11	.3665	.1335	1.51	.4345	.0655
1.12	.3686	.1314	1.52	.4357	.0643
1.13	.3708	.1292	1.53	.437	.063
1.14	.3729	.1271	1.54	.4382	.0618
1.15	.3749	.1251	1.55	.4394	.0606
1.16	.377	.123	1.56	.4406	.0594
1.17	.379	.121	1.57	.4418	.0582
1.18	.381	.119	1.58	.4429	.0571
1.19	.383	.117	1.59	.4441	.0559

continued on next page

Table 12.1 continued from previous page

$z$	$\mu$ to $z$	beyond $z$	$z$	$\mu$ to $z$	beyond $z$
1.6	.4452	.0548	2	.4772	.0228
1.61	.4463	.0537	2.01	.4778	.0222
1.62	.4474	.0526	2.02	.4783	.0217
1.63	.4484	.0516	2.03	.4788	.0212
1.64	.4495	.0505	2.04	.4793	.0207
1.65	.4505	.0495	2.05	.4798	.0202
1.66	.4515	.0485	2.06	.4803	.0197
1.67	.4525	.0475	2.07	.4808	.0192
1.68	.4535	.0465	2.08	.4812	.0188
1.69	.4545	.0455	2.09	.4817	.0183
1.7	.4554	.0446	2.1	.4821	.0179
1.71	.4564	.0436	2.11	.4826	.0174
1.72	.4573	.0427	2.12	.483	.017
1.73	.4582	.0418	2.13	.4834	.0166
1.74	.4591	.0409	2.14	.4838	.0162
1.75	.4599	.0401	2.15	.4842	.0158
1.76	.4608	.0392	2.16	.4846	.0154
1.77	.4616	.0384	2.17	.485	.015
1.78	.4625	.0375	2.18	.4854	.0146
1.79	.4633	.0367	2.19	.4857	.0143
1.8	.4641	.0359	2.2	.4861	.0139
1.81	.4649	.0351	2.21	.4864	.0136
1.82	.4656	.0344	2.22	.4868	.0132
1.83	.4664	.0336	2.23	.4871	.0129
1.84	.4671	.0329	2.24	.4875	.0125
1.85	.4678	.0322	2.25	.4878	.0122
1.86	.4686	.0314	2.26	.4881	.0119
1.87	.4693	.0307	2.27	.4884	.0116
1.88	.4699	.0301	2.28	.4887	.0113
1.89	.4706	.0294	2.29	.489	.011
1.9	.4713	.0287	2.3	.4893	.0107
1.91	.4719	.0281	2.31	.4896	.0104
1.92	.4726	.0274	2.32	.4898	.0102
1.93	.4732	.0268	2.33	.4901	.0099
1.94	.4738	.0262	2.34	.4904	.0096
1.95	.4744	.0256	2.35	.4906	.0094
1.96	.475	.025	2.36	.4909	.0091
1.97	.4756	.0244	2.37	.4911	.0089
1.98	.4761	.0239	2.38	.4913	.0087
1.99	.4767	.0233	2.39	.4916	.0084

continued on next page

Table 12.1 continued from previous page

$z$	$\mu$ to $z$	beyond $z$	$z$	$\mu$ to $z$	beyond $z$
2.4	.4918	.0082	2.8	.4974	.0026
2.41	.492	.008	2.81	.4975	.0025
2.42	.4922	.0078	2.82	.4976	.0024
2.43	.4925	.0075	2.83	.4977	.0023
2.44	.4927	.0073	2.84	.4977	.0023
2.45	.4929	.0071	2.85	.4978	.0022
2.46	.4931	.0069	2.86	.4979	.0021
2.47	.4932	.0068	2.87	.4979	.0021
2.48	.4934	.0066	2.88	.498	.002
2.49	.4936	.0064	2.89	.4981	.0019
2.5	.4938	.0062	2.9	.4981	.0019
2.51	.494	.006	2.91	.4982	.0018
2.52	.4941	.0059	2.92	.4982	.0018
2.53	.4943	.0057	2.93	.4983	.0017
2.54	.4945	.0055	2.94	.4984	.0016
2.55	.4946	.0054	2.95	.4984	.0016
2.56	.4948	.0052	2.96	.4985	.0015
2.57	.4949	.0051	2.97	.4985	.0015
2.58	.4951	.0049	2.98	.4986	.0014
2.59	.4952	.0048	2.99	.4986	.0014
2.6	.4953	.0047	3	.4987	.0013
2.61	.4955	.0045	3.01	.4987	.0013
2.62	.4956	.0044	3.02	.4987	.0013
2.63	.4957	.0043	3.03	.4988	.0012
2.64	.4959	.0041	3.04	.4988	.0012
2.65	.496	.004	3.05	.4989	.0011
2.66	.4961	.0039	3.06	.4989	.0011
2.67	.4962	.0038	3.07	.4989	.0011
2.68	.4963	.0037	3.08	.499	.001
2.69	.4964	.0036	3.09	.499	.001
2.7	.4965	.0035	3.1	.499	.001
2.71	.4966	.0034	3.11	.4991	.0009
2.72	.4967	.0033	3.12	.4991	.0009
2.73	.4968	.0032	3.13	.4991	.0009
2.74	.4969	.0031	3.14	.4992	.0008
2.75	.497	.003	3.15	.4992	.0008
2.76	.4971	.0029	3.16	.4992	.0008
2.77	.4972	.0028	3.17	.4992	.0008
2.78	.4973	.0027	3.18	.4993	.0007
2.79	.4974	.0026	3.19	.4993	.0007

continued on next page

Table 12.1 continued from previous page

$z$	$\mu$ to $z$	beyond $z$	$z$	$\mu$ to $z$	beyond $z$
3.2	.4993	.0007	3.6	.4998	.0002
3.21	.4993	.0007	3.61	.4998	.0002
3.22	.4994	.0006	3.62	.4999	.0001
3.23	.4994	.0006	3.63	.4999	.0001
3.24	.4994	.0006	3.64	.4999	.0001
3.25	.4994	.0006	3.65	.4999	.0001
3.26	.4994	.0006	3.66	.4999	.0001
3.27	.4995	.0005	3.67	.4999	.0001
3.28	.4995	.0005	3.68	.4999	.0001
3.29	.4995	.0005	3.69	.4999	.0001
3.3	.4995	.0005	3.7	.4999	.0001
3.31	.4995	.0005	3.71	.4999	.0001
3.32	.4995	.0005	3.72	.4999	.0001
3.33	.4996	.0004	3.73	.4999	.0001
3.34	.4996	.0004	3.74	.4999	.0001
3.35	.4996	.0004	3.75	.4999	.0001
3.36	.4996	.0004	3.76	.4999	.0001
3.37	.4996	.0004	3.77	.4999	.0001
3.38	.4996	.0004	3.78	.4999	.0001
3.39	.4997	.0003	3.79	.4999	.0001
3.4	.4997	.0003	3.8	.4999	.0001
3.41	.4997	.0003	3.81	.4999	.0001
3.42	.4997	.0003	3.82	.4999	.0001
3.43	.4997	.0003	3.83	.4999	.0001
3.44	.4997	.0003	3.84	.4999	.0001
3.45	.4997	.0003	3.85	.4999	.0001
3.46	.4997	.0003	3.86	.4999	.0001
3.47	.4997	.0003	3.87	.4999	.0001
3.48	.4997	.0003	3.88	.4999	.0001
3.49	.4998	.0002	3.89	.4999	.0001
3.5	.4998	.0002	3.9	.5	0
3.51	.4998	.0002	3.91	.5	0
3.52	.4998	.0002	3.92	.5	0
3.53	.4998	.0002	3.93	.5	0
3.54	.4998	.0002	3.94	.5	0
3.55	.4998	.0002	3.95	.5	0
3.56	.4998	.0002	3.96	.5	0
3.57	.4998	.0002	3.97	.5	0
3.58	.4998	.0002	3.98	.5	0
3.59	.4998	.0002	3.99	.5	0

Table 12.1 also can be used to determine the probabilities for ranges not explicitly listed in the table. For example, what proportion of the scores from a normal distribution are between 1 and 2 standard deviations from the mean? As 47.72% of the scores are between the mean and 2 standard deviations from the mean, and 34.13% are between the mean and 1 standard deviation, then  $47.72 - 34.13 = 13.59\%$  of the scores are between 1 and 2 standard deviations from the mean.

### 12.1.1 The Shah (1985) approximation

The values in Table 12.1 should be accurate to the number of decimal places displayed. They were computed *not* by performing all the relevant integrations (which would be tedious to compute, to say the least), but by using one of the many *approximation algorithms* that have been developed over the years.<sup>2</sup> Some of these are quite complex, and accordingly quite accurate. However, even quite simple algorithms can deliver reasonably accurate results. Probably the simplest while still being reasonably accurate of these is the one proposed by Shah (1985). To two decimals of precision, the approximation has 3 rules:

1. For  $0 \leq z \leq 2.2$ , the proportion of area between the mean and  $z$  is given as  $z(4.4 - z)/10$ .
2. For  $2.2 < z < 2.6$ , the proportion is .49,
3. and for  $z \geq 2.6$ , the proportion is .5.

For example, for  $z = 1$ , the first rule is applied, yielding  $(1)(4.4 - 1)/10 = .34$ , as compared to the .3413 of Table 12.1. According to Shah, comparing the simple approximation values to the actual areas yielded a maximum absolute error of .0052, or roughly only 1/2%.

## 12.2 The Normal Approximation to the Binomial

As the normal curve is expressed in terms of standard deviations from the mean (as in Table 12.1), to use it to approximate large- $n$  binomial distributions requires that we express the binomial statistic (i.e.,  $\sum x$ ) as a standard score. To do that, we need to know the mean and standard deviation of the binomial distribution of  $\sum x$ . The mean is simple enough. If there are  $n$  trials, and the probability of observing the positive event on any give trial is  $p$ , the *expected value* or mean is simply  $np$ . The variance is similarly determined. If  $p$  is close to 1, then most events will be positive, and there will be little variability, whereas, the closer it is to .5, the more variable it will be, reaching a maximum when  $p = .5$ . And the

---

<sup>2</sup>The values shown in Table 12.1 were computed using the *Statistics Toolbox* in *MATLAB*, a computer software package. It is not known what particular approximation algorithm is used by this software.

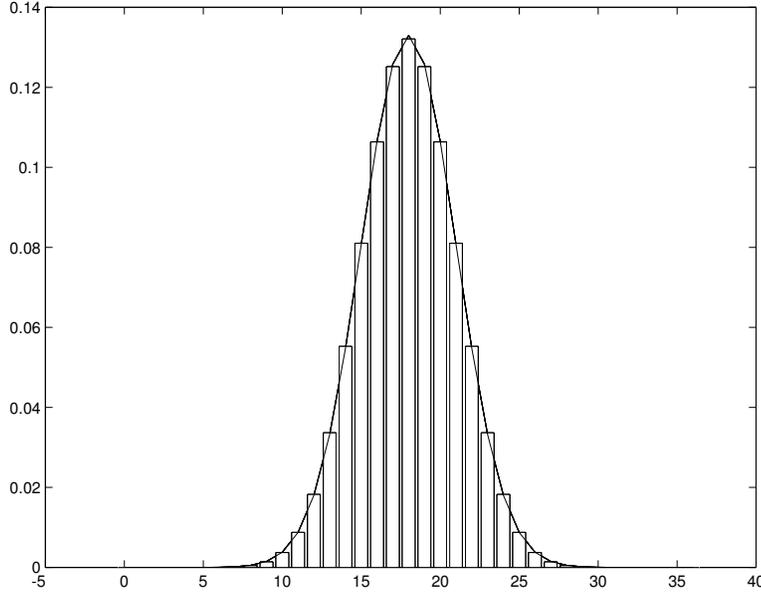


Figure 12.2: The binomial distribution for  $n = 36, p = .5$  overlaid with a normal approximation with  $\mu = 18, \sigma = 3$ .

more events we sum over, the larger the variability will be. Thus, the variance of the sum should be a direct function of  $n$  and an inverse function of  $p > .5$ , or  $np(1-p)$ , and the standard deviation will be  $\sqrt{np(1-p)}$ .

Thus, to convert any value from a binomial distribution to a standard score:

$$z = \frac{x - np}{\sqrt{np(1-p)}}$$

For example, 24 positive events in 36 Bernoulli trials (with  $p = .5$ ) expressed as a standard score is  $(24 - 36 * .5) / \sqrt{36 * .5 * (.5)} = (24 - 18) / 3 = 2$ . Using Table 12.1, we need only then convert the obtained standard score to a tail probability to assess the statistical significance of the obtained binomial result.<sup>3</sup> For the current example, that would be  $\alpha = .0228$ , one-tailed test, or  $\alpha = .0456$ , two-tailed test. The exact value is  $.0326$ , one-tailed test, emphasizing that the normal approximation *is*, after all, only an approximation. The approximation improves, of course, as  $n$  gets larger. It can also be improved by adding a “correction for continuity”—an adjustment that takes into account the fact that

<sup>3</sup>To use the binomial distribution directly to obtain the tail probability, we would have to compute the sum of

$$\binom{36}{24} p^{24} q^{36-24} + \binom{36}{25} p^{25} q^{36-25} + \dots + \binom{36}{36} p^{36} q^0$$

which could take awhile.

the binomial is a discrete distribution but the normal used to approximate it is a continuous distribution.<sup>4</sup>

A plot of the binomial for  $n = 36, p = .5$  is shown in Figure 12.2, overlaid with a plot of the normal distribution with  $\mu = 18$  and  $\sigma = 3$ . As can be seen, the normal *is* a very good fit, except for the discreteness of the bars of the binomial. In practice, though, the error is usually not that serious, especially as  $n$  increases. Below about  $n = 30$ , however, the normal approximation to the binomial is considered too coarse to use; this is especially true if  $p$  differs from .5.

## 12.3 The Normal Approximation to the Distributions of Sums From Any Distributions

The normal approximation to the binomial represents only a special case of a much more general result—and one known as the *central limit theorem*. Recall that the binomial distribution comes about as the distribution of the sum of  $n$  Bernoulli trials, scored as 0 and 1. That is, we define a statistic, say  $X$ , to equal the sum over each of the  $n$  Bernoulli trials, each of which could have been either a 0 or 1. We then determine the *relative frequency distribution* of the statistic  $X$  over the range from each of the  $n$  trials coming up 0 through half zeroes, half ones, to every trial coming up 1.

What if on each trial instead of sampling from a binomial (0/1), we sampled from a different distribution, and possibly a different distribution on each trial? On trial 1, we sample from distribution  $X_1$  with mean  $= \mu_1$  and variance  $= \sigma_1^2$ , on trial 2, from distribution  $X_2$  with mean  $= \mu_2$  and variance  $= \sigma_2^2$ , and so on up to trial  $n$ , and distribution  $X_n$  with mean  $= \mu_n$  and variance  $= \sigma_n^2$ . As with the binomial, let us define the statistic  $X = X_1 + X_2 + \cdots + X_n$ . As long as each of the trials is *independent* of the others, then under certain, very general conditions (which almost always hold) and large  $n$ ,

$$Z_n = \frac{X - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \quad (12.3)$$

is also distributed as a standard normal distribution.

### 12.3.1 Sums of Normal Distributions

If each of the distributions contributing to the sum is itself *normal*, then equation 12.3 will be distributed as a standard normal regardless of the size of  $n$ . To see how useful this result can be, consider the following example. You sample, say, 9 scores from a distribution known to be normal with  $\mu = 50$ , and  $\sigma = 6$ . Assuming that the sample is a *random sample* (which means each score was selected independently of the others), what is the probability that the mean of

<sup>4</sup>One such correction is to subtract .5 from the discrete value at the bottom of the range and add .5 to the discrete value at the top of the range.

the 9 scores,  $\bar{X}$ , would be greater than 54? Means are just sums corrected for  $n$ , so because  $\mu$  and  $\sigma$  are the same for each score (i.e., independent trial), equation 12.3 could be re-written in terms of means as:

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \quad (12.4)$$

Substituting the values from our example,

$$\begin{aligned} Z_{\bar{X}} &= \frac{54 - 50}{\sqrt{\frac{6^2}{9}}} \\ &= \frac{4}{\frac{6}{\sqrt{9}}} \\ &= 2 \end{aligned}$$

we find from Table 12.1 that  $\Phi(z > 2) = .0228$ .

### 12.3.2 The Standard Error of the Mean

The denominator in equation 12.4 is known as the *standard error of the mean*, and is usually symbolized as:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} \quad (12.5)$$

### 12.3.3 Sums of Nonnormal Distributions

Of course, the principle advantage of the central limit theorem is that we can work with sums, and thereby means, from distributions other than normal distributions, as long as  $n$  is large enough. “Large enough” here, as with the binomial, is generally accepted to be  $n$  greater than about 30. For example, imagine we sample 36 (i.e.,  $> 30$ ) scores from a distribution of either unknown character or known not to be normal, say a *uniform distribution* over the range of 10 to 30. The mean of a uniform distribution, as one might expect, is given by the sum of the endpoints divided by two, so, in this case,  $\mu = (30 + 10)/2 = 20$ . The standard deviation of a uniform distribution is given the square-root of the squared difference of the endpoints divided by 12,<sup>5</sup> so, in this case,  $\sigma = \sqrt{\frac{(30-10)^2}{12}} = 20/\sqrt{12} = 5.77$ . With those parameters in hand, what is the probability of observing a sample mean of our 36 scores,  $\bar{X}$ , greater than, say, 22, assuming (as always) that it is a random sample? Substituting these values into equation 12.4,

$$Z_{\bar{X}} = \frac{22 - 20}{\sqrt{\frac{5.77^2}{36}}}$$

---

<sup>5</sup>Yes, 12, for any and all uniform distributions; the reason for the denominator of 12 is left as an exercise for the reader.

$$\begin{aligned}
 &= \frac{2}{\frac{5.77}{\sqrt{36}}} \\
 &= 2.08
 \end{aligned}$$

we find from Table 12.1 that  $\Phi(z > 2.08) = .0188$ .

### 12.3.4 A Confidence Trick

Jerzy Neyman (1894–1981), like Fisher, was one of the luminaries in the development of statistics subsequent to Karl Pearson. He published extensively with Karl Pearson’s son, Egon Pearson; indeed, Neyman and Pearson were responsible for the development of an important alternative to Fisher’s significance testing approach (involving both alternative and null hypotheses—indeed, they introduced the concepts of null and alternative hypotheses, Type I and Type II errors, and Power), now often attributed to Fisher! Of his many brilliant statistical innovations, Neyman’s proposal of the “confidence interval” is probably one of the most heralded, if still one of the most controversial.

Neyman promoted the view that all significance tests were tests of population parameters. Thus, the test in the just preceding section (section 12.3.3), for Neyman (and most other statisticians to this day, incidentally) is seen as a test that the population mean,  $\mu$ , from which the sample of 36 scores allegedly has been drawn is actually equal to 20. According to Fisher’s significance test, having rejected the model, we have no idea what the actual population mean might be, only that it is unlikely to be equal to 20 if the data were a sample from the specified uniform distribution. Neyman argued differently. Assuming the model is otherwise correct, what is the likely range of values the population mean could take? Rather than centre on the hypothesised population mean,  $\mu = 20$ , and computing how different (plus or minus) our sample mean would have to be to be considered significantly different (at whatever alpha-level we have chosen), why not centre on the sample mean, and compute what range of population means,  $\mu$ , would be consistent (within the same alpha-level considerations) with that sample mean?

Again, consider the previous example. We have a sample of 36 scores from a hypothesised uniform distribution with  $u = 20$  and a standard error of the mean of

$$\sigma_{\bar{X}} = \frac{5.77}{\sqrt{36}}$$

With  $\alpha = .05$ , as we have already shown, our sample mean would have to be *at least* plus or minus 1.96 of these standard errors of the mean from the population mean to be declared significantly different (at the *alpha* = .05 level) from the hypothesised mean of 20 (assuming the rest of the model is true). That means that, assuming the model is true, 95% of the time we would observe sample means within 1.96 standard errors of the mean of  $\mu = 20$ . But, as Neyman noted, that must also mean, assuming the model is true except for the value of the parameter  $\mu$ , that the actual value of the population mean,  $\mu$ , must also be

within 1.96 standard errors of the mean of the sample mean 95% of the time. That is,

$$-1.96\sigma_{\bar{X}} + \bar{X} \leq \mu \leq 1.96\sigma_{\bar{X}} + \bar{X}$$

is the 95% *confidence interval*. These are the same intervals that accompany election and other poll results in which you are told, for example, that the “poll percentages are reliable plus or minus 3% 19 times out of 20”. For the current example, the 95% confidence interval is

$$-1.96 \frac{5.77}{\sqrt{36}} + 22 \leq \mu \leq 1.96 \frac{5.77}{\sqrt{36}} + 22 = 20.12 \leq \mu \leq 22.96$$

(Note that the lower bound is greater than our hypothesised mean, which we already knew because we computed the sample mean of 22 to be significantly different from a population mean of 20.)

Neyman referred to these ranges as “confidence intervals” because it is not at all clear exactly what the 95% actually refers to. From a frequentist probabilist perspective it can be interpreted to mean that if you repeated the sampling many, many times, all other things being equal, 95% of those times the interval would encompass the actual value of  $\mu$  (given, again, that the model is true except for the parameter,  $\mu$ ). That is, you are “95% confident” that the interval overlaps the actual value of the population parameter. But that isn’t saying much; in particular, it *doesn’t* say that there is a 95% probability that the population mean falls within the computed bounds. And, despite its many proponents who argue for confidence intervals as a superior replacement for Fisherian significance tests, it really says no more than the original significance test did. You can, of course, compute 99% confidence intervals, and 99.9% intervals (and 90% and 80% intervals), but they provide no more information (just in a different form) than would the significance test at the corresponding alpha-level. Fisher, for one, never bought into them. And, inductively, they add nothing to what the significance test already provided: if the significance test was significant, then the corresponding confidence interval will *not* overlap the hypothesised mean; otherwise, it will.

Subsequent chapters present similarly-constructed inferential statistics, all of which can equally, using exactly the same logic, have confidence intervals computed for the putative “tested” parameter. If the conditions are appropriate for a significance test, then the computed confidence intervals are merely redundant. If more is assumed (or obtained), such as true *random sampling* (i.e., so that the random sampling serves a role beyond that of providing the independence assumed by significance testing), then confidence intervals can be serve as valuable tools in the evaluation of estimates of population parameters from sample statistics. However, in that case the goals are fundamentally different.

### 12.3.5 Why it’s called the “Normal Distribution”

There is no reason to limit our sampling to samples from a single distribution of scores. Suppose we are interested in the end result of a process that is the linear

combination (i.e., sum) of a large number of other, independent processes. Some possible examples could be your final score on a large, multiple choice exam, your height, and your speed on an open highway. In each case, the particular value can be seen to be the result (i.e., the sum) of many independent positive and negative factors. If so, then according to the central limit theorem, the distribution of these values (over individuals or over multiple trials for one individual) should be approximately normal, especially if the number of independent factors,  $n$ , is large.

Many variables or distributions of human interest (e.g., height, weight, IQ, GRE scores, mean recognition on a memory test, number of errors in solving a maze) can be seen to be result of many independent factors; according to the central limit theorem, such distributions should also be approximately normal. And they are. It is for this reason—that the distributions of many traits and variables should be approximately normal—that it is referred to as the “normal distribution”—the one to “expect”, unless something unusual is going on.

Of course, many of the variables we record are not, in the main, the result of many independent factors. In fact, often it is our hope, especially in science, that the variables we measure reflect the operation of only a single factor so that the variation we observe in this variable is principally due to differences in the value of the underlying factor. But even in those situations for which, with careful control, we have eliminated a major influence of other factors, it is still the case that the act of measurement itself is often subject to the influence of multiple, albeit slight, factors. Thus, even though the bulk of the variance in the scores is not a function of multiple factors, and hence, unlikely to be normally distributed, small errors in the measurement of the “true” value *are* likely to be distributed normally. In fact, the normal distribution via the central limit theorem was first proposed as a model of error by Laplace in 1810, and was explicitly used as model of measurement error by the astronomer F. W. Bessel in 1818.

Many current descriptions of the use of the normal distribution in statistics adhere to this idea of an observed value as consisting of a true value plus normally distributed error with a mean of zero around it. More generally, is the idea that the core of statistical data analysis consists of the following statement:

$$\text{data} = \text{model} + \text{error}$$

That is, data (observations) have two independent sources: the model of the data (such as a linear regression equation) and error in the fit of the model. The error term is assumed (via the central limit theorem) to be normally distributed with a mean of zero (the standard deviation of the error is estimated from the data). Why a mean of zero? Because we can always add a simple additive constant to the model (especially simple linear models) to take into account any non-zero average deviation of the error term. Implicit in this approach, however, is the more direct question of “model fitting”—that is, how well does some proposed (theoretical or otherwise) distribution of parameters match the statistics of the distribution actually obtained? Once again, Karl Pearson was the principle source of these techniques, and they are the subject of the next chapter.

## 12.4 Questions

1. Thirty-six scores, randomly-sampled from a population of scores with  $\sigma = 18$ , are found to have a mean of 10. What is the probability that the population mean ( $\mu$ ) is less than 7?
2. You are given a coin and told that either it is a fair coin or that it is a coin biased to come up “heads” 80% of the time. To figure out which it is, you toss the coin 36 times and count the number of heads. Using the notion that the coin is fair as your null hypothesis, and setting  $\alpha = .10$ ,
  - (a) what is the probability of a type II error?
  - (b) what would it be if you tossed the coin 64 times?
3. Despite the fact that grades are known not to be normally distributed, comparisons of mean grades (say, between one class and another) often are done using the normal distribution as the theoretical (null hypothesis) model of how such means would be distributed. Explain this apparent contradiction.
4. A coin is tossed 64 times and 28 heads are observed. Based on the evidence, is it likely (say,  $\alpha = .05$ ) to be a fair coin?
5. From a normal distribution of scores with a mean of 20 and a standard deviation of 5, 25 scores are selected at random and their mean calculated. What is the probability that the mean is less than 18.5?
6. A local dairy sells butter in 454 gram (1 pound, for those wedded to imperial measures) packages. As a concerned consumer, you randomly sample 36 of these packages (all that your research budget can afford), record their weights as measured on a home scale, and compute a mean of 450 grams. As this is below the stated weight, you contact the dairy and are informed by a spokesperson that your mean represents simply the typical variability associated with the manufacturing process. The spokesperson notes further that the standard deviation for all packages of butter produced by the dairy is 12 grams. Should you believe the claims of the spokesperson? Show why or why not.
7. The weights of Loonies (Canada’s \$1.00 coins) in circulation are normally distributed with a mean of 700 and a standard deviation of 10 (measured in hundredths of a gram). Banks receive coins rolled in paper tubes. You have been given 100 loonies so rolled, but suspect that some (i.e., at least one) of these loonies are actually slugs (slugs are either too heavy or too light relative to actual loonies). Describe how you could confirm or reject your suspicion to a reasonable level of certainty (say,  $\alpha = .05$ ) without having to unroll the coins. State the null and alternative hypotheses, the test statistic and how you would calculate it, what value(s) of the test statistic would be critical, etc.

8. When there is no fire, the number of smoke particles in the air is distributed normally with a mean of 500 parts per billion (PPB) and a standard deviation of 200 ppb. In the first minute of a fire, the number of smoke particles is likewise normally distributed, but with a mean of 550 ppb and a standard deviation of 100 ppb. For a particular smoke detector:
  - (a) what criterion will detect 88.49% of the fires in the first minute?
  - (b) what will be the false-alarm rate for the criterion in a.?
  - (c) with the criterion set to produce only 6.68% false-alarms, what will be the probability of a type II error?
9. With some obvious exceptions (hee hee!), women's hair at the U. of L. is generally longer than men's. Assume that the length of women's hair at the U. of L. is normally distributed with  $\mu = 44$  and  $\sigma = 8$  cm. For men at the U. of L., assume that the length of hair is also normally-distributed, but with  $\mu = 34$  and  $\sigma = 16$  cm. Four heads of hair are sampled at random from *one* of the two sexes and the mean length is calculated. Using a decision criterion of 40 cm for discriminating from which of the two sexes the hair sample was drawn,
  - (a) what is the probability that an all male sample will be correctly labeled as such?
  - (b) what is the probability that an all female sample will be *mislabeled*?
10. Although there probably are some students who do actually like sadistics (sic), common sense says that in the general population, the proportion is at most 1/3. Of 36 of Vokey and Allen's former sadistics students, 16 claim to like sadistics. Test the hypothesis that these 36 people are a random sample of the general population. Be sure to articulate the null hypothesis, and any assumptions necessary for drawing the conclusion in question.
11. From a set of scores uniformly distributed between 0.0 and  $\sqrt{12}(8)$ , 64 scores are sampled at random. What is the probability that the mean of the sample is greater than 12?

## Chapter 13

# The $\chi^2$ Distribution

Chapter 12 introduced the central limit theorem, and discussed its use in providing for the normal approximation to the binomial distribution. This chapter introduces another application of both the theorem and the normal approximation, enroute to another continuous, theoretical distribution known as the chi (pronounced *ky*—rhymes with “sky”) square distribution, symbolized as  $\chi^2$ .

Consider again the binomial situation of Chapter 12, only this time imagine that we were examining the results of tossing a coin, say, 64 times. According to our model of a fair coin (and tossing procedure), we would expect  $\mu = .5(64) = 32$  heads, on average. Suppose, however, that we observed only  $x = 24$  heads. Is this result sufficient evidence to reject our model of a fair coin?

Converting the observed frequency of 24 to a standardized score as in Chapter 12, yields:

$$\begin{aligned} z &= \frac{x - \mu}{\sqrt{n(p)(1 - p)}} \\ &= \frac{24 - 32}{\sqrt{64(.5)(1 - .5)}} \\ &= -2 \end{aligned} \tag{13.1}$$

Using the normal approximation to the binomial, observing 24 heads (or fewer) in 64 tosses of a fair coin should occur less than requisite 5% of the time (2.28% to be exact), so we should reject the model.

### 13.1 Chi square and the goodness of fit

There is another, much more general way of arriving at the same conclusion, all due to the work, once again, of Karl Pearson. First, we need to compute a new statistic. Let's call this new statistic  $z^2$ . And let's compute it as the ratio of

the square of the difference between what was observed and what was expected (i.e., theorised) over what was expected for each of the two possible outcomes, and then sum them. That is, where  $O_i$  and  $E_i$  refer to observed and expected frequencies, respectively:

$$\begin{aligned}
 z^2 &= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \\
 &= \frac{(24 - 32)^2}{32} + \frac{(40 - 32)^2}{32} \\
 &= \frac{64}{32} + \frac{64}{32} \\
 &= 4
 \end{aligned} \tag{13.2}$$

Note that  $z^2$  is a measure of how well the *expected* frequencies,  $E_1$  and  $E_2$ , fit the *observed* frequencies,  $O_1$  and  $O_2$ . The closer the expected frequencies are to the obtained frequencies, the smaller  $z^2$  will be, reaching zero when they are equal. Hence, if we knew the distribution of the statistic  $z^2$  we could compare our computed value to that distribution, and, hence, determine the likelihood or relative frequency of such a result (or a more extreme result), assuming that the obtained value of  $z^2$  was just a random value from that distribution.

There is a family of theoretical distributions, called the chi square distributions, that are defined as the sum of independent, squared scores from standard normal distributions for  $n > 1$ , and as a squared normal distribution if only one is involved. That is, where  $x$  is a score from any normal distribution, and  $n$  indexes the number of independent scores summed over, the corresponding value from the requisite  $\chi^2$  distribution is given as:

$$\chi_n^2 = \sum_{i=1}^n \left( \frac{x_i - \mu_i}{\sigma_i} \right)^2 \tag{13.3}$$

Shown in Figure 13.1 are four chi square distributions from the family of such distributions. Chi square distributions are distinguished by the number of independent, squared, standard normal distributions summed over to produce them. This parameter is referred as the number of “degrees of freedom”—independent sources—that went in to producing the distribution. This concept of “degrees of freedom” will become clearer subsequently (and, not so incidentally, was the source of one of Karl Pearson’s biggest failures, and the principle source of his life-long feud with Fisher).

As surprising as it may appear, the computation of  $z^2$  in equation 13.2 is arithmetically just an alternative method of computing  $z$  in equation 13.1, except that it computes the *square* of  $z$  (note for our example values that 4 is the square of -2, and this will be the case for all frequencies of  $O_i$  and  $E_i$  and, hence,  $z$  and  $z^2$ ). Now, if it is reasonable to invoke the normal approximation to the binomial for our example in the application of equation 13.1, then it is equally reasonable to assume it in the arithmetically related case of  $z^2$  in equation 13.2; that is, that  $z^2$  is the corresponding *square* of a score from a standard normal

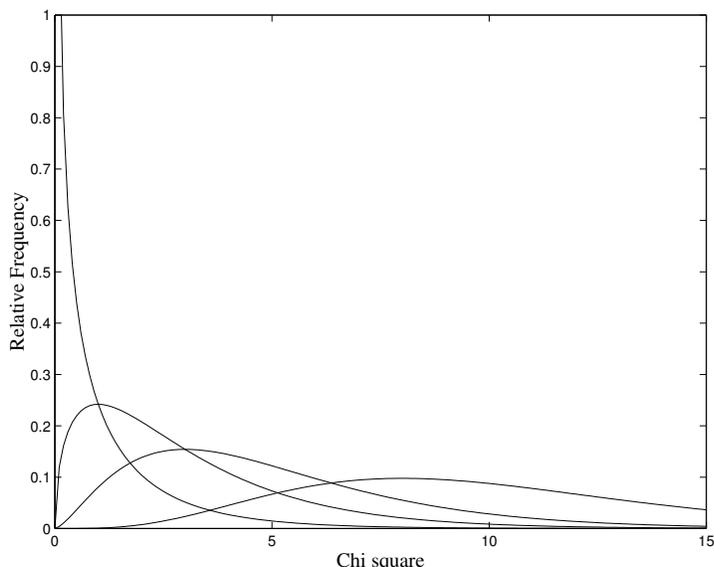


Figure 13.1: Relative frequency (probability density) as a function of values of  $\chi^2$ . A plot of the chi square distribution for 1 (the backward j-curve), and (moving left to right), 3, 5, and 10 degrees of freedom.

distribution, and, hence, is distributed as  $\chi^2$  with 1 degree of freedom (i.e., is a score from the distribution  $\chi_1^2$ ).

Table 13.1 lists the critical tail values of  $\chi^2$  for different  $\alpha$  levels as a function of degrees of freedom ( $df$ ) (or which particular  $\chi^2$  distribution is being referred to). According to Table 13.1, a value of  $\chi^2$  with 1  $df$  (see line 1 of Table 13.1) corresponding to the  $z^2$  of 4 from our example should occur somewhere between 2% and 5% of the time (i.e.,  $\chi^2 = 4$  is between 3.84 and 5.41 in Table 13.1). The precise value, not shown in Table 13.1, is 2.28% of the time—exactly as we determined with the normal approximation to the binomial. In fact, you could expand on the values listed in line 1 of Table 13.1 simply by squaring the value of  $z$  from Table 12.1 in Chapter 12 corresponding to the  $p$ -value of interest.

Apart from providing another method (perhaps easier, perhaps not) of computing the normal approximation to the binomial, what's the point?

The point, as Pearson showed, is that the principle extends beyond that of the binomial situation. Consider a *polynomial* situation, say, one with five categories: Liberal, Conservatives, NDP, Greens, and Other, representing national political party preferences. Suppose the results of a recent poll of a statistics class were that of the 60 students, 20 preferred the Liberals, another 20 preferred the NDP, 10 preferred the Conservatives, 5 preferred the Greens, with the remaining 5 distributed over the various parties of “Other”. Suppose further that a national poll taken at roughly the same time resulted in the corresponding percentages

df	Level of $\alpha$								
	0.25	0.2	0.15	0.1	0.05	0.02	0.01	.001	.0001
1	1.32	1.64	2.07	2.71	3.84	5.41	6.63	10.83	15.14
2	2.77	3.22	3.79	4.61	5.99	7.82	9.21	13.82	18.42
3	4.11	4.64	5.32	6.25	7.81	9.84	11.34	16.27	21.11
4	5.39	5.99	6.74	7.78	9.49	11.67	13.28	18.47	23.51
5	6.63	7.29	8.12	9.24	11.07	13.39	15.09	20.52	25.74
6	7.84	8.56	9.45	10.64	12.59	15.03	16.81	22.46	27.86
7	9.04	9.8	10.75	12.02	14.07	16.62	18.48	24.32	29.88
8	10.22	11.03	12.03	13.36	15.51	18.17	20.09	26.12	31.83
9	11.39	12.24	13.29	14.68	16.92	19.68	21.67	27.88	33.72
10	12.55	13.44	14.53	15.99	18.31	21.16	23.21	29.59	35.56
11	13.7	14.63	15.77	17.27	19.68	22.62	24.73	31.26	37.37
12	14.85	15.81	16.99	18.55	21.03	24.05	26.22	32.91	39.13
13	15.98	16.98	18.2	19.81	22.36	25.47	27.69	34.53	40.87
14	17.12	18.15	19.41	21.06	23.68	26.87	29.14	36.12	42.58
15	18.25	19.31	20.6	22.31	25	28.26	30.58	37.7	44.26
16	19.37	20.47	21.79	23.54	26.3	29.63	32	39.25	45.92
17	20.49	21.61	22.98	24.77	27.59	31	33.41	40.79	47.57
18	21.6	22.76	24.16	25.99	28.87	32.35	34.81	42.31	49.19
19	22.72	23.9	25.33	27.2	30.14	33.69	36.19	43.82	50.8
20	23.83	25.04	26.5	28.41	31.41	35.02	37.57	45.31	52.39
21	24.93	26.17	27.66	29.62	32.67	36.34	38.93	46.8	53.96
22	26.04	27.3	28.82	30.81	33.92	37.66	40.29	48.27	55.52
23	27.14	28.43	29.98	32.01	35.17	38.97	41.64	49.73	57.07
24	28.24	29.55	31.13	33.2	36.42	40.27	42.98	51.18	58.61
25	29.34	30.68	32.28	34.38	37.65	41.57	44.31	52.62	60.14
26	30.43	31.79	33.43	35.56	38.89	42.86	45.64	54.05	61.66
27	31.53	32.91	34.57	36.74	40.11	44.14	46.96	55.48	63.16
28	32.62	34.03	35.72	37.92	41.34	45.42	48.28	56.89	64.66
29	33.71	35.14	36.85	39.09	42.56	46.69	49.59	58.3	66.15
30	34.8	36.25	37.99	40.26	43.77	47.96	50.89	59.7	67.63

Table 13.1: Critical tail values of the  $\chi^2$ -distribution as a function of degrees of freedom ( $df$ ) and selected levels of  $\alpha$ .

Liberal	NDP	Conservatives	Green	Other	Total ( $n$ )
20 (30)	20 (9)	10 (9)	5 (6)	5 (6)	60

Table 13.2: The observed and expected (in parentheses) frequencies of national political party preferences for the example polynomial situation discussed in the text.

of 50%, 15%, 15%, 10%, and 10%, respectively. Are the results of the statistics class significantly different from that *expected* on the basis of the national poll?

According to the national poll, 50% or 30 of the 60 statistics students would be expected to have preferred the Liberals, 9 to have preferred the NDP, 9 to have preferred the Conservatives, 6 to have preferred the Greens, and the remaining 6 to have preferred Other. These observed and expected values are shown in Table 13.2. Extending equation 13.2, where  $k$  refers to the number of cells, we get:

$$\begin{aligned}
 z^2 &= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} && (13.4) \\
 &= \frac{(20 - 30)^2}{30} + \frac{(20 - 9)^2}{9} + \frac{(10 - 9)^2}{9} + \frac{(5 - 6)^2}{6} + \frac{(5 - 6)^2}{6} \\
 &= 17.22
 \end{aligned}$$

For the 5 cells of observed frequencies in Table 13.2 (with their total), only 4 contribute independent information to the computation of  $z^2$ , because if you know the frequencies in any 4 of the cells, the fifth can be determined by subtraction from the total. This example demonstrates the meaning of degrees of freedom: in a set of 5 frequencies only 4 are free to vary because the fifth is fixed by the total frequency. In general, with  $n$  frequencies, only  $n - 1$  are free to vary. Hence, for such tables, there are  $n - 1$  degrees of freedom. So, entering Table 13.1 with  $5 - 1 = 4$  *df* indicates that obtaining a  $z^2$  value of 17.22 or larger would occur somewhere between 0.1% and 1% of the time *if* the observed frequencies were drawn at random from a model with those expected frequencies. Because the  $p$ -value is less than .05, the model is rejected, and we conclude that the results from the statistics class are significantly different from what would be expected on the basis of the national poll.

Despite the fact that the polynomial situation is not binomial, use of the  $\chi^2$  distribution still requires the same conditions to be met for it to be a meaningful model: the categories must be exhaustive, mutually exclusive (any given case can fall into one and only category), and assignment of a case to any given category must occur independently of the assignment of any other case to its category—this last requirement to ensure the independence required of the normal approximation. Thus, for example, combining the 10 responses from you and from each of your friends on a recent multiple-choice test to assess, say, whether the distribution of responses into correct and incorrect categories was significantly different from chance would most likely be inappropriate because

	Coke	Pepsi	
Females	32	28	60
Males	8	32	40
	40	60	100

Table 13.3: An example of a 2 x 2 contingency table depicting the cross-classification of cola preference by sex.

the responses from any one of you are unlikely to be truly independent of one another. Finally, remember that the use of the  $\chi^2$  distribution here is just an *approximation* to the actual discrete distribution of frequencies: if the overall  $n$  is too small or the expected frequency for any cell is too low, the approximation breaks down. A rule of thumb here is that the expected frequency for every cell should be greater than about 5.

## 13.2 $\chi^2$ tests of independence

### 13.2.1 2 x 2 contingency tables

The *contingency table* shown in Table 13.3 depicts the cross-classification of 100 students by sex (female or male) and their cola preference (Coke or Pepsi). There are at least three questions one might want to ask of these data. The first two, concerning whether the *marginal* distribution of cola preference or the marginal distribution of the sexes differ from some specified distributions (say, 50/50 split), are the subject of the previous section 13.1. The third question, though, is different: one might be interested in whether the distribution of cola preference varies as a function of sex; that is, whether cola preference is *contingent* upon the sex of the individual stating the preference—whether cola preference is *independent* of sex for this table. Once again, Karl Pearson led the way, developing tests of independence.

Chi square tests of independence do not differ fundamentally from tests of goodness of fit. The computation and underlying logic of  $\chi^2$ , for example, are exactly the same. They differ, however, in the purpose of the test, and in how the expected frequencies are obtained. Tests of goodness of fit are intended to assess whether or the extent to which a particular theoretical distribution of expected frequencies fits the obtained frequencies. Tests of independence, on the other hand, are used to assess whether or the extent to which two cross-classifications of the frequencies are *independent* of each other. To do so, the expected frequencies rather than being given *a priori* by some theoretical distribution are derived instead from this notion of independence. That is, they are the cell frequencies that would be obtained *if* the cross-classifications, such as sex and cola preference in Table 13.3, were indeed completely independent.

To be independent means that one should not be able to predict the cola preference from knowledge about sex (or vice versa). Shown in Table 13.4 is the same cross-classification of cola preference and sex as that shown in

	Coke	Pepsi	
Females	24	36	60
Males	16	24	40
	40	60	100

Table 13.4: An example of a 2 x 2 contingency table in which the cross-classifications of cola preference and sex are independent.

Table 13.3, except that the four cell frequencies have been adjusted so that cola preference is now independent of sex. Note that the *marginal totals* for both cola preference and sex are identical to those of the original table; the adjustment for independence does not affect these totals. Instead, the interior cell frequencies are adjusted so that the *proportion* of females who prefer Coke is the same as the proportion males who do (or, identically, that the proportion of those with a Cola preference that are female is the same as that with a Pepsi preference).

So, for example,  $24/60 = .40$  or 40% of the females in Table 13.4 prefer Coke, as do  $16/40 = .4$  or 40% of the males. Similarly,  $24/40 = .60$  or 60% of the Coke drinkers are female, as are  $36/60 = .60$  or 60% of the Pepsi drinkers. Note that these proportions are identical to those of the corresponding marginals (i.e., 40% of the 100 people prefer Coke, and 60% of the 100 are females), which, in fact, is how they were computed. To compute the independence or expected frequency for the cell corresponding to females with a Coke preference, for example, just multiply the proportion of the *total* frequency that preferred Coke (.4) by the marginal frequency of females (60) [or, equivalently, multiply the proportion of the *total* frequency that were female (.6) by the marginal frequency of those that prefer Coke (40)] to get 24. In general, to compute the expected frequency for a cell in row  $i$ , column  $j$  given independence, where  $m_i$  and  $m_j$  are the corresponding row and column marginal totals, and  $n$  is the total frequency:

$$E_{ij} = \frac{m_i m_j}{n} \quad (13.5)$$

### Degrees of freedom

For 2 x 2 contingency tables, only one of the four expected frequencies needs to be computed in this way; the remainder can be obtained by subtraction from the corresponding marginal totals to complete the table. What this result means is that for the assessment of independence in a 2 x 2 table, only one cell frequency is free to vary; once one cell is known, the others are automatically determined. Hence, for the assessment of independence in a 2 x 2 table, there is only 1 degree of freedom.

Once all the expected frequencies have been determined,  $z^2$  is computed in same way as before (see equation 13.4):

$$z^2 = \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i}$$

$$\begin{aligned}
&= \frac{(32 - 24)^2}{24} + \frac{(28 - 36)^2}{36} + \frac{(8 - 16)^2}{16} + \frac{(32 - 24)^2}{24} \\
&= 11.11
\end{aligned}$$

Comparing this result to the  $\chi^2$  distribution with 1 *df* in Table 13.1, reveals that a value of  $\chi^2$  at least this large will occur only between 0.01% and 0.1% of time. Hence, the result is significantly different from that expected if cola preference and sex actually were independent. As can be seen in Table 13.3, on average, females prefer Coke and males prefer Pepsi.

### A measure of association

To say that cola preference and sex are *not* independent is to say that they are *correlated* or *associated* with one another. So, the first question one might ask following rejection of independence is, if cola preference and sex are associated, just how associated are they? A common measure of association for 2 x 2 contingency tables is called the *phi* (pronounced *fee*) coefficient, which is a simple function of the approximated value of  $\chi^2$  ( $z^2$  in our case) and  $n$  (the total frequency):

$$\phi = \sqrt{\frac{z^2}{n}} \quad (13.6)$$

The  $\phi$  coefficient varies between zero and 1, much like the absolute value of the Pearson product-moment correlation coefficient. In fact, it *is* the Pearson product-moment correlation coefficient. Exactly the same result would be obtained if the values of 0 and 1 were assigned to males and females, respectively, the values of 0 and 1 were assigned to the Coke and Pepsi preferences, respectively, and the techniques of Chapter 7 were applied to the 100 cases in Table 13.3. For our example case,

$$\begin{aligned}
\phi &= \sqrt{\frac{\chi^2}{n}} \\
&= \sqrt{\frac{11.11}{100}} \\
&= 0.33
\end{aligned}$$

That is, cola preference and sex are moderately (but significantly!) correlated or associated in Table 13.3.

### 13.2.2 r x c contingency tables

As with the tests of goodness of fit that extended beyond the binomial case, tests of independence are not limited to 2 x 2 contingency tables, but can have any number  $r$  of rows and any number  $c$  of columns. Calculation of the expected values proceeds in the same way as with 2 x 2 tables, as given by equation 13.5, as does the computation of  $z^2$ , as given by equation 13.4.

What does change is the computation of the degrees of freedom. For greater than  $2 \times 2$  tables, where  $r$  is the number of rows and  $c$  is the number of columns, the degrees of freedom is found by:

$$df = (r - 1)(c - 1) \quad (13.7)$$

So, for example, for a contingency table with 4 rows and 6 columns, the degrees of freedom would be:

$$\begin{aligned} df &= (r - 1)(c - 1) \\ &= (4 - 1)(6 - 1) \\ &= 15 \end{aligned}$$

However,  $df$  here is still just the number of cells that have to be filled in before the remainder can be determined by subtraction from the marginal totals.

#### A measure of association for $r \times c$ tables

The  $\phi$ -coefficient in equation 13.6 can't be used for  $r \times c$  tables because it is based on the two cross-classifications having exactly two levels each. A related measure—a generalisation of the  $2 \times 2$  *phi* coefficient—is available for  $r \times c$  tables, called Cramér's  $\phi$ . Where  $z^2$  is our approximated value of  $\chi^2$ ,  $n$  is the total frequency for the table, and  $k$  is the *smaller* of  $r$  and  $c$ :

$$\phi_C = \sqrt{\frac{z^2}{n(k - 1)}} \quad (13.8)$$

Cramér's  $\phi$  is similar to  $\phi$ : it varies between 0 and 1, and higher values are indicative of greater levels of association. But, unlike  $\phi$ , it is *not* a Pearson product-moment correlation coefficient.

### 13.3 Questions

1. The Consumers association of Canada (CAC) decides to investigate the claims of a new patent cold remedy. From 25 years of investigations conducted at the British Cold institute, it is known that for people who take no medication at all, one week from the onset of cold symptoms 30% show improvement, 45% show no change, and the remainder are worse. The results for a sample of 50 people who used the new cold medicine were that 10 people improved, 25 showed no change, and the remainder got worse. Based on these results, what should the CAC conclude about the new cold remedy, and why?
2. Although there probably are some students who do actually like sadistics (sic), common sense says that in the general population, the proportion is at most  $1/3$ . Of 36 of Vokey and Allen's former sadistics students, 16 claim

to like sadistics. Test the hypothesis that these 36 people are a random sample of the general population. Be sure to articulate null and alternative hypotheses, and any assumptions necessary for drawing the conclusion in question.

3. A study was conducted to determine whether there is a relationship between socioeconomic status (high/low) and political party support. The following results were obtained:

SES	NDP	LIB	CONS	Other
high	21	24	46	7
low	28	10	6	6

- (a) Is there a relationship between socioeconomic status (high/low) and political party support? Of what magnitude?
- (b) Is there sufficient evidence to support the claim that the two status classes were sampled equally?
- (c) Ignoring status (and the category “Other”), are the political parties represented equally?
4. Two successive flips of a fair coin is called a trial. 100 trials are run with a particular coin; on 22 of the trials, the coin comes up “heads” both times; on 60 of the trials the coin comes up once a “head” and once a “tail”; and on the remaining trials, it comes up “tails” for both flips. Is this sufficient evidence ( $\alpha = .05$ ) to reject the notion that the coin (and the flipping process) is a fair one?
5. From a distribution of scores claimed to be normal with  $\mu = 50$  and  $\sigma = 10$ , you randomly take 50 samples of 25 scores each and calculate the mean of each sample. 30 of these means are between 48 and 52, 5 are greater than 52, and the remainder are less than 48. Are these data sufficient to reject the claim at the  $\alpha = .05$  level?
6. An intrepid student decides to complete her honours research project at the “end-of-term” student pub. Ensnared as the bartender, she uses as her sample the first 60 students to walk up to the bar, and for each one records the individual’s sex and whether or not he or she orders a beer. She constructs the following contingency table:

	Male	Female
Beer	20	15
No beer	5	20

- (a) Based on these data, is there a significant relationship between sex and the type of drink? Of what magnitude?

- (b) In random samples of the student population, beer-drinkers generally out-number non-beer drinkers 2 to 1. Is there sufficient evidence to reject the hypothesis that these 60 students are a random sample of the student population?



## Chapter 14

# The $t$ Distribution

All of the foregoing was very much dependent on the fact that in each case the population variance,  $\sigma^2$ , was known. But what if  $\sigma^2$  is *not* known? This was the problem faced by W. S. Gosset, a chemist at Guinness breweries (see Figure 14.1), in the first few decades of the twentieth century. One approach, the one current at the time, would be to use *very* large samples, so that the variance of the sample would differ hardly at all from the population or process sampled from, and simply substitute the sample variance for the population variance in equation 12.4. But as reasonable as this solution appears to be (assuming the sample size really is *large*), it would not work for Gosset. His problem was to work with samples taken from tanks of beer to determine, for example, whether the concentration of alcohol was different from its optimum at that point in the brewing process.<sup>1</sup> The use of large samples would use up the product he was testing! He needed and developed a *small-sample* approach.

### 14.1 Student's $t$ -test

He started with the idea that because of the central limit theorem, it often would be reasonable for many variables and processes to assume that the *population* distribution of these variables and processes would be normal with some unknown variance. Now, as a *random sample* from that population, the variance of the sample,  $S^2$ , could be used to *estimate* that unknown population variance. It wouldn't be exact, but it would appear to be better than nothing. Unfortunately, as it turns out,  $S^2$  is a *biased* estimate of  $\sigma^2$ —it is more likely to be *too small* as to be *too large*. That is, in addition to the error introduced by the fact that

---

<sup>1</sup>His initial problem was actually to develop a method to assess the yeast concentration in vats of yeast. Precise volumes of yeast of equally precise concentrations according to a (still) secret formula had to be added to the mix of hops and barley to produce the beer for which Guinness was justly so famous. Too much or too little would ruin the beer. But, the alive yeast were constantly reproducing in the vats, so the concentrations fluctuated daily, and had to be assessed just before being added to the next batch of beer.



Figure 14.1: W. S. Gosset (“Student”)

the sample statistics are merely *estimates* of the population parameters,  $S^2$  in general would tend to be too low as an estimate of  $\sigma^2$ .

To see why this bias is so, consider as an example the *population* consisting of the three scores  $\{1, 2, 3\}$ . Clearly, the *mean* of this population,  $\mu$ , is equal to  $(1 + 2 + 3)/3 = 2$ , and the variance of this population,  $\sigma^2$ , is equal to  $((1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2)/3 = 2/3 = 0.67$ . Now consider all the possible samples of three numbers from this population. Sampling with replacement, that means we can have samples such as  $\{1, 1, 1\}$ , through  $\{1, 2, 3\}$ , to  $\{3, 3, 3\}$ . Table 14.1 lists each of the possible  $3 \times 3 \times 3 = 27$  samples, along with their corresponding means and variances. Note that in Table 14.1, although the *mean* of the sample means *equals* the population mean (of 2.0)—and, hence, is an *unbiased* estimator of the population mean, the mean of the sample variances (0.44) is *less than* the population variance (of 0.67)—on average, it *underestimates* the population variance, and, hence, is *biased* estimator of it.

The appropriate correction for this bias in the estimate is to compute the estimate as the sum of squared deviations from the mean divided by  $n - 1$ , rather

Sample	Sample Mean	Variance: $S^2$	Est. Variance: $\hat{\sigma}^2$
1 1 1	1	0	0
1 1 2	1.33	0.22	0.33
1 1 3	1.67	0.89	1.33
1 2 1	1.33	0.22	0.33
1 2 2	1.67	0.22	0.33
1 2 3	2	0.67	1
1 3 1	1.67	0.89	1.33
1 3 2	2	0.67	1
1 3 3	2.33	0.89	1.33
2 1 1	1.33	0.22	0.33
2 1 2	1.67	0.22	0.33
2 1 3	2	0.67	1
2 2 1	1.67	0.22	0.33
2 2 2	2	0	0
2 2 3	2.33	0.22	0.33
2 3 1	2	0.67	1
2 3 2	2.33	0.22	0.33
2 3 3	2.67	0.22	0.33
3 1 1	1.67	0.89	1.33
3 1 2	2	0.67	1
3 1 3	2.33	0.89	1.33
3 2 1	2	0.67	1
3 2 2	2.33	0.22	0.33
3 2 3	2.67	0.22	0.33
3 3 1	2.33	0.89	1.33
3 3 2	2.67	0.22	0.33
3 3 3	3	0	0
Means	2	0.44	0.67

Table 14.1: An example of bias in using sample variances ( $S^2$ ) as estimates of population variance. The means, sample variances,  $S^2$ , and unbiased estimates of population variance,  $\hat{\sigma}^2$ , of the 27 possible samples from the population  $\{1, 2, 3\}$ . Note that the *average* of the sample variances (0.44) is *less than* the population variance (0.67), but that the *average* of the population estimates equals the population variance.

than  $n$ :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Why  $n - 1$ ? The reason is that variance is the average squared deviation from the mean; it is based on deviation scores from which the mean has been removed—the deviation scores are constrained to have a mean of zero. This constraint is equivalent to knowing one of the scores contributing to the sum, and that score, therefore, is no longer free to vary. For example, if you know that the sum of two scores is equal to 5, then once you find the value of just one of the scores, the value of the other is determined because they must sum to 5. The *variance* possible for these scores, then, is in this sense determined by the variability of just one of them, rather than both. Thus, if we want the average of the variability possible of a set of scores, the average should be based on the number contributing to it, which is one less than the number of scores. This estimate,  $\hat{\sigma}^2$ , is no more likely to be too small as to be too large, as shown in the last column of Table 14.1, and, hence, is an *unbiased* estimate of the population variance.

But it is still an estimate, which means that a different sample would likely give rise to a different estimate. That is, these *estimates* are themselves distributed.<sup>2</sup> Thus, even if we substituted these now unbiased estimated values into equation 12.4, the resulting distribution of  $Z$  would *not* be normal in general, because  $Z$  is now the distribution of the ratio of one random distribution (the assumed normal) and another random distribution (that of the estimated variance), the result of which is known mathematically *not* to be normal, although it converges on the normal as  $n$  gets large (principally because, as  $n$  gets larger, the denominator converges on a constant,  $\sigma^2/\sqrt{n}$ , and is no longer distributed). In fact, although the shapes are similar (i.e., they are all symmetrical, roughly triangular-shaped distributions similar to the normal), the exact shape of the distribution and, hence, the proportions of scores that fall into different ranges, changes as a function of how many scores went into computing the estimated variance. This number, the denominator of the formula for computing the estimated variance,  $n - 1$ , is referred to as the *degrees of freedom* of the estimate.

This family of related distributions is called the  $t$ -distributions, and Gosset worked out what the exact distributions were for each of the small samples he would likely have to work with, such as sample sizes of less than 10 or 20. A selection of tail values of the  $t$ -distributions for many of these small samples is shown in Table 14.2, as a function of degrees of freedom, and for both one- and two-tailed tests. He published these ideas and distributions in a series of articles under the pseudonym “Student”, both because the Guinness brewing company had an explicit policy forbidding its employees to publish methods related to the brewing of beer after an unfortunate incident involving another employee in which a trade secret had been so revealed, and because he fancied himself at

---

<sup>2</sup>Given the assumption of normality for the population, they are in fact distributed as a chi-square ( $\chi^2$ ) distribution with  $n - 1$  degrees of freedom. See Chapter 13.

best as just a student of the then burgeoning work of such statistical luminaries as Galton, Pearson, and Fisher.

### 14.1.1 *t*-test for one sample mean

As an example of using the *t*-distributions, consider that you have just taken 9 sampled scores from some process known or thought to give rise to a normal distribution of values, with a mean  $= \mu = 20$  and an unknown standard deviation  $= \sigma$ , as long as the process has not been disturbed. You suspect, though, that the process is degenerating, giving rise to lower values than it should. Imagine that the mean and variance of your sample were  $\bar{X} = 18$  and  $S^2 = 8$ . The sample mean being less than the hypothesized mean of the population is consistent with your suspicions, but, of course, could have arisen simply as a function of the sampling of the 9 scores from the population (i.e., some of the 9-score samples from the population could have means of 18, or even less). To test whether your sample mean of 18 is deviant enough from the hypothesized value of 20 to confirm your suspicions that the process has degenerated, you conduct a *t*-test.<sup>3</sup> First, the sample variance,  $S^2 = 8$ , is converted to an estimate of the population variance, as follows. To recover the sample sum-of-squares, the sample variance is multiplied by  $n$ . Then, the sum-of-squares is divided by the degrees of freedom,  $n - 1$ , to yield an unbiased estimate of the population variance,  $\hat{\sigma}^2$ . For our example, then, we get

$$\begin{aligned}\hat{\sigma}^2 &= \frac{S^2 n}{n - 1} \\ &= \frac{(8)(9)}{(9 - 1)} \\ &= 9\end{aligned}$$

That value is the *estimated population variance*, but what we want is an *estimate of the standard deviation of sample means* from such a population, which we obtain, as we did earlier, by dividing the estimated population variance by  $n$ , and taking the square-root, producing an *estimated standard error of the mean*:

$$\hat{\sigma}_{\bar{X}} = \sqrt{\frac{\hat{\sigma}_{\bar{X}}^2}{n}} \quad (14.1)$$

Formally, the *t*-test for testing a single mean against its hypothesized population value is defined as a modification of equation 12.4:

$$t_{df} = \frac{\bar{X} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \quad (14.2)$$

---

<sup>3</sup>Gosset used the symbol *z*; subsequent authors, keen to distinguish Gosset's test from the *z*-test, introduced the *t* symbol. Why *t* was chosen, rather than, say, *g* (for Gosset) or *s* (for Student) is not known to the authors.

df	Level of $\alpha$ for a one-tailed test					
	.2	.10	.05	.025	.01	.001
	Level of $\alpha$ for a two-tailed test					
	.4	.2	.1	.05	.02	.002
1	1.3764	3.0777	6.3138	12.7062	31.8205	318.3088
2	1.0607	1.8856	2.9200	4.3027	6.9646	22.3271
3	0.9785	1.6377	2.3534	3.1824	4.5407	10.2145
4	0.9410	1.5332	2.1318	2.7764	3.7469	7.1732
5	0.9195	1.4759	2.0150	2.5706	3.3649	5.8934
6	0.9057	1.4398	1.9432	2.4469	3.1427	5.2076
7	0.8960	1.4149	1.8946	2.3646	2.9980	4.7853
8	0.8889	1.3968	1.8595	2.3060	2.8965	4.5008
9	0.8834	1.3830	1.8331	2.2622	2.8214	4.2968
10	0.8791	1.3722	1.8125	2.2281	2.7638	4.1437
11	0.8755	1.3634	1.7959	2.2010	2.7181	4.0247
12	0.8726	1.3562	1.7823	2.1788	2.6810	3.9296
13	0.8702	1.3502	1.7709	2.1604	2.6503	3.8520
14	0.8681	1.3450	1.7613	2.1448	2.6245	3.7874
15	0.8662	1.3406	1.7531	2.1314	2.6025	3.7328
16	0.8647	1.3368	1.7459	2.1199	2.5835	3.6862
17	0.8633	1.3334	1.7396	2.1098	2.5669	3.6458
18	0.8620	1.3304	1.7341	2.1009	2.5524	3.6105
19	0.8610	1.3277	1.7291	2.0930	2.5395	3.5794
20	0.8600	1.3253	1.7247	2.0860	2.5280	3.5518
21	0.8591	1.3232	1.7207	2.0796	2.5176	3.5272
22	0.8583	1.3212	1.7171	2.0739	2.5083	3.5050
23	0.8575	1.3195	1.7139	2.0687	2.4999	3.4850
24	0.8569	1.3178	1.7109	2.0639	2.4922	3.4668
25	0.8562	1.3163	1.7081	2.0595	2.4851	3.4502
26	0.8557	1.3150	1.7056	2.0555	2.4786	3.4350
27	0.8551	1.3137	1.7033	2.0518	2.4727	3.4210
28	0.8546	1.3125	1.7011	2.0484	2.4671	3.4082
29	0.8542	1.3114	1.6991	2.0452	2.4620	3.3962
30	0.8538	1.3104	1.6973	2.0423	2.4573	3.3852
40	0.8507	1.3031	1.6839	2.0211	2.4233	3.3069
60	0.8477	1.2958	1.6706	2.0003	2.3901	3.2317
80	0.8461	1.2922	1.6641	1.9901	2.3739	3.1953
100	0.8452	1.2901	1.6602	1.9840	2.3642	3.1737
$\infty$	0.8416	1.2816	1.6449	1.9600	2.3264	3.0902

Table 14.2: Critical values of the  $t$ -distribution.

So, for the current case with  $9 - 1 = 8$  degrees of freedom, or *df*:

$$\begin{aligned} t_8 &= \frac{18 - 20}{\sqrt{\frac{9}{9}}} \\ &= -2 \end{aligned}$$

Looking up the absolute value of  $t$ ,  $|t|$ , in Table 14.2 (because, as with the standard normal table, only the positive tail of the symmetric  $t$ -distribution is tabulated) for 8 *df*, we find that to obtain significance at an  $\alpha$ -level of .05 for a one-tailed test (which is appropriate in the case as either no difference or a sample mean greater than the hypothesized population mean would have contradicted our idea that the process in question was degenerating toward lower values) requires a  $t$ -value of 1.86 or greater. Our absolute value of 2 exceeds this value, so we declare the result significant at the  $\alpha = .05$  level, rejecting the model with  $\mu = 20$  as the explanation for our obtained result.

### 14.1.2 $t$ -test for two *dependent* sample means

Here is another example of exactly the same test involving, superficially, the means from two dependent or *correlated* samples.<sup>4</sup> To show the parallels with the foregoing, suppose, again, that we have 9 samples, but this time they come in *pairs*, such as scores on test 1 and matching scores on test 2 for each of 9 students (or vats of beer, in Gosset's case). Suppose that we suspect that scores over the two tests are dropping—that test 2 was on average more difficult than test 1. Suppose the 9 students had a mean of 20 on test 1, but only a mean of 18 on test 2, consistent with our hypothesis, but, of course, it is possible such a difference could arise simply by randomly-sampling the scores (or, more accurately in this case, randomly-sampling the *differences*). *Subtracting* each student's test 1 score from his or her test 2 score yields a sample of 9 *difference* scores with a mean,  $\bar{D} = -2$ , and, say, a variance of  $S_D^2 = 8$ . If the scores are *not* dropping, it seems reasonable to assume that the *population* of such difference scores should give rise to a normal distribution of values, with a mean  $= \mu = 0$  and an unknown standard deviation  $= \sigma$ . We are now back to exactly the case described earlier, except that for our observed mean difference of  $\bar{D} = -2$ , our null hypothesis mean difference is  $\mu = 0$ , or no difference between test 1 and test 2. According to equation 14.2:

$$\begin{aligned} t_8 &= \frac{-2 - 0}{\sqrt{\frac{9}{9}}} \\ &= -2 \end{aligned}$$

Our absolute value of 2 exceeds the critical  $t_8$ -value of 1.86 or greater, so we declare the result significant at the  $\alpha = .05$  level, rejecting the model with  $\mu = 0$  or no difference between the two tests as the explanation for our obtained result.

<sup>4</sup>Also often referred to as the  $t$ -test for *correlated* means or samples, the  $t$ -test for *matched samples*, the *dependent*  $t$ -test, and so on. Regardless of the label, they are all just variants of the  $t$ -test for one sample mean.

## 14.2 $t$ -test for two *independent* sample means

Probably the most commonly used statistical test is the  $t$ -test for the difference between two sample means, such as the mean of an experimental group versus that of a control group, the mean for a sample of males versus that of a corresponding sample of females, and so on. The principal idea here is a simple extension of the one-sample  $t$ -test just described in section 14.1.

The model or null hypothesis in this case is that we have two, independent random samples of scores from the same normally-distributed population of scores, or, equivalently, that the samples are drawn at random from normally-distributed populations of scores having the same mean and variance. However, because they are just samples, both their means and their variances are likely to differ somewhat from the population parameters, and from one another. Given the model, we can compute how likely or unlikely such differences are, and reject the model if the differences we actually obtained would be too unlikely if it were true.

Because both groups of scores are assumed to be random samples from the same population of scores, we could use either one to estimate the population variance, as in section 14.1. However, as they are both supposed to be random samples of the same population, we should get an even better estimate if we combine or *pool* them because, as in section 14.1, the accuracy of the estimate improves as sample size, or  $n$  increases. Because the samples are not necessarily the same size (i.e.,  $n_1 \neq n_2$ ), we want to weight (or bias) the pooled estimate in favour of the one with the larger sample size, as follows. With  $S_1^2$  as the variance of sample 1, and  $S_2^2$  as the variance of sample 2, we first recover the sum-of-squares from each sample, add them together to get the total sum-of-squares, and then divide by the total degrees of freedom:

$$\hat{\sigma}_{pooled}^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$\hat{\sigma}_{pooled}^2$  is now our best estimate of the variance of this hypothetical population of scores from which our two samples of scores were presumed to have been drawn. In section 14.1, the denominator of the equation 14.2, the square-root of the estimated population standard deviation divided by  $n$ , is the *estimated standard error of the mean* or estimated standard deviation of means of samples of size  $n$ . We need something similar here, although as we are interested in the difference between two sample means, we want the *estimated standard error of the difference* between sample means. Even though we are subtracting the means from one another, the variability of each mean contributes to the variability of the difference. Thus, we want to *add* our best estimate of the variability of one mean to the corresponding best estimate of the variability of the other mean to get the best estimate of the variance of the difference between the means, and then take the square-root to get the estimated standard error of the difference

between two means (for those sample sizes):

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\hat{\sigma}_{pooled}^2}{n_1} + \frac{\hat{\sigma}_{pooled}^2}{n_2}} \quad (14.3)$$

The  $t$ -test is now just the ratio of the difference between the two sample means over the estimated standard error of such differences:

$$t_{df} = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}} \quad (14.4)$$

where the degrees of freedom are, as before, the  $df$  that produced the estimated variance,  $\hat{\sigma}_{pooled}^2$ , in this case,  $n_1 + n_2 - 2$ . As a demonstration of the  $t$ -test for two sample means, let us return to canonical example of section 10.1.3 of chapter 10 in which females had the scores  $\{1, 2, 3, 5\}$  and males had the scores  $\{4, 6, 7, 8\}$ . To make the example more concrete, imagine these scores were errors on a sadistics (*sic*) test. The *null hypothesis model* is that these errors are distributed normally, and the distribution of scores is identical for males and females as we have selected them (i.e., the scores for our groups represent independent random samples from this distribution of scores).

The mean for the males is 6.25 errors, with the mean number of errors for the females as 2.75, for a difference between the two means of 3.5. The sample variance in each case is  $S^2 = 2.1875$ , so the pooled estimated variance is equal to:

$$\begin{aligned} \hat{\sigma}_{pooled}^2 &= \frac{4(2.1875) + 4(2.1875)}{(4-1) + (4-1)} \\ &= 2.92 \end{aligned}$$

and the standard error of the estimate of the difference between the means is equal to:

$$\begin{aligned} \hat{\sigma}_{\bar{X}_1 - \bar{X}_2} &= \sqrt{\frac{2.92}{4} + \frac{2.92}{4}} \\ &= 1.21 \end{aligned}$$

Finally, the  $t$ -ratio with  $(4-1) + (4-1) = 6$  degrees of freedom for a difference between the means of 3.5 is equal to:

$$\begin{aligned} t_6 &= \frac{3.5}{1.21} \\ &= 2.90 \end{aligned}$$

Looking at the  $t$ -table with 6 degrees of freedom, yields a critical value of 2.4469 at  $\alpha = .05$ , two-tailed test. As 2.90 exceeds that value, we reject the model, and declare that the males scored significantly more errors on average than did the females.

### 14.2.1 Relationship between $t$ and $r_{pb}$

Remember from Chapter 7, section 7.3.2, on the point-biserial correlation coefficient, that one can compute  $r_{pb}$  in exactly the same situation as that for the  $t$ -test for two independent sample means; indeed,  $t$  and  $r_{pb}$  are essentially the same statistic—one is just an algebraic transformation of the other:<sup>5</sup>

$$r_{pb} = \pm \sqrt{\frac{t_{n-2}^2}{t_{n-2}^2 + (n-2)}}$$

$$t_{n-2} = \frac{r_{pb}\sqrt{n-2}}{\sqrt{1-r_{pb}^2}}$$

Thus, the  $t$ -test is *also* a test of the *significance* of the  $r_{pb}$  statistic—if the corresponding  $t$  is significant, then the correlation between group membership and the dependent measure is *significantly* different from zero. By the same token,  $r_{pb}$  is the corresponding measure of the *strength of relationship* between group membership and the dependent measure for the  $t$ -test. Thus, for the example in the previous section (14.2):

$$r_{pb} = \pm \sqrt{\frac{2.9^2}{2.9^2 + 6}}$$

$$r_{pb} = \pm 0.76$$

That is, group membership (sex) is highly correlated with errors, or, equivalently, that  $.76^2 = 57.76\%$  of the variance in the errors can be “accounted for” by group membership.

## 14.3 Questions

1. Sixty-four people are asked to measure the width of a large room. The mean measurement is 9.78 metres with a variance of 63 squared-centimetres. Assuming that the measurement errors are random (and normally distributed), the true width of the room is unlikely (i.e.,  $\alpha = .05$ ) to be less than \_\_\_ or greater than \_\_\_.
2. From a normally-distributed population of scores with  $\mu = 40$ , a random sample of 17 scores is taken, and the sample mean calculated. The standard deviation for the sample is 10, and the mean is found to be *greater than* the population mean. If the probability of observing a sample mean of that magnitude (or greater) is equal to .10, what is the sample mean?
3. A set of 65 scores, randomly sampled from a normal distribution, has a mean of 30 and a variance of 1600. Test the notion that population from which this sample was drawn has a mean of 40.

<sup>5</sup>A derivation that we will leave as an exercise for the reader ...

4. From a normally-distributed population of scores with a mean of  $\mu$ , 9 scores are sampled at random. The mean and standard deviation for this sample of 9 scores are found to be 12 and 4, respectively.  $\mu$  is unlikely ( $\alpha = .05$ ) to be less than \_\_\_ or greater than \_\_\_.
5. Over the years, students not seeking extra assistance in statistics (sic) from the instructor have obtained a mean grade of 70 in the course. A group of 36 students who did seek extra assistance obtained a mean grade of 67 with a standard deviation (for the group) of 9. Perform a statistical test to determine whether the mean for the group is significantly different ( $\alpha = .05$ ) from that of the population of students who did not seek extra assistance.
6. On five different trips by automobile over the same route between your home in Lethbridge and a friend's home near Calgary, you record the following odometer readings (in kilometres): 257, 255, 259, 258, and 261. Assuming that measurement errors of this kind would be normally-distributed, is there sufficient evidence to reject the belief that the distance between your home and your friend's over this route is 260 km?
7. Many people claim that the television volume surges during commercials. To test this notion, you measure the average volume during 36 regular programs and the corresponding commercials that occur within them, producing 36 pairs of program-commercial volumes. On average, the commercials are 8 decibels louder than the programs, and the sum of squared differences is equal to 16329.6. Is this sufficient evidence to substantiate the claim?
8. The manufacturer of the *Lamborghini Countach* recently claimed in a prestigious magazine that the sports car was able to accelerate from 0-60 in 4.54 seconds. In my previous life as a test driver for this company (we all can dream, can't we?), I drove 25 of these fine automobiles and found that the mean latency in acceleration from 0-60 was 4.84 seconds, with a standard deviation (for the sample) of .35 seconds. Is my result statistically inconsistent with the claim in the magazine?



## Chapter 15

# The General Logic of ANOVA

The statistical procedure outlined here is known as “analysis of variance”, or more commonly by its acronym, ANOVA; it has become the most widely used statistical technique in biobehavioural research. Although randomisation testing (e.g., permutation and exchangeability tests) may be used for completely-randomized designs (see Chapter 10),<sup>1</sup> the discussion here will focus on the normal curve method (nominally a *random-sampling* procedure<sup>2</sup>), as developed by the British statistician, Sir Ronald A. Fisher

### 15.1 An example

For the example data shown in Table 15.1, there are three scores per each of three independent groups. According to the usual null hypothesis, the scores in each group comprise a random sample of scores from a *normally-distributed* population of scores.<sup>3</sup> That is, the groups of scores are assumed to have been sampled from normally-distributed populations having equal means and variances

---

<sup>1</sup>Randomisation testing also can be used when random assignment has not been used, in which case the null hypothesis is simply that the data are distributed as one would expect from random assignment—that is, there is no systematic arrangement to the data.

<sup>2</sup>When subjects have not been randomly sampled, but have been randomly assigned, ANOVA and the normal curve model may be used to estimate the  $p$ -values expected with the randomization test. It is this ability of ANOVA to approximate the randomization test that was responsible at least in part for the ready acceptance and use of ANOVA for experimental research in the days before computing equipment was readily available to perform the more appropriate, but computationally-intensive randomisation tests. Inertia, ignorance, or convention probably account why it rather than randomisation testing is still used in that way.

<sup>3</sup>Note that it is the theoretical distribution from which the scores themselves are being drawn that is assumed to be normal, and not the sampling distribution. As we will see, the sampling distribution of the statistic of interest here is definitely not normal.

Group		
1	2	3
-2	-1	0
-1	0	1
0	1	2
$\bar{X}_1 = -1$	$\bar{X}_2 = 0 = \bar{X}_G$	$\bar{X}_3 = 1$

Table 15.1: Hypothetical values for three observations per each of three groups.

(or, equivalently, from the *same* population). Actually, the analysis of variance (or *F*-test), as with Student's *t*-test (see Chapter 14, is fairly robust with respect to violations of the assumptions of normality and homogeneity of variance,<sup>4</sup> so the primary claim is that of equality of means; the alternative hypothesis, then, is that at least one of the population means is different from the others.

### 15.1.1 Fundamental ANOVA equation

Analysis of variance is so-called because it tests the null hypothesis of equal means by *testing the equality of two estimates of the population variance of the scores*. The essential logic of this procedure stems from the following tautology. For any score in any group (i.e., score *j* in group *i*), it is obvious that:

$$X_{ij} - \bar{X}_G = X_{ij} - \bar{X}_i + \bar{X}_i - \bar{X}_G$$

That is, the deviation of any score from the grand mean of all the scores is equal to the deviation of that score from its group mean plus the deviation of its group mean from the grand mean. Or, to put it more simply, the *total* deviation of a score is equal to its *within* group deviation plus its *between* group deviation.

Now, if we square and sum these deviations over all scores:<sup>5</sup>

$$\sum_{ij} (X_{ij} - \bar{X}_G)^2 = \sum_{ij} (X_{ij} - \bar{X}_i)^2 + \sum_{ij} (\bar{X}_i - \bar{X}_G)^2 + 2 \sum_{ij} (X_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}_G)$$

For any group, the group mean minus the grand mean is a constant, and the sum of deviations around the group mean is always zero, so the last term is always zero, yielding:

$$\sum_{ij} (X_{ij} - \bar{X}_G)^2 = \sum_{ij} (X_{ij} - \bar{X}_i)^2 + \sum_{ij} (\bar{X}_i - \bar{X}_G)^2$$

or

<sup>4</sup>Which is to say that the *p*-values derived from its use are not strongly affected by such violations, as long as the violations are not too extreme.

<sup>5</sup>Because, it will be remembered, the sum of deviations from the mean over all scores is always zero; so, we always square the deviations before summing—producing a *sum of squares*, or *SS* (see Chapter 2).

	$X_{ij}$	$\bar{X}_i$	$\bar{X}_G$	$(X_{ij} - \bar{X}_G)^2$	$(X_{ij} - \bar{X}_i)^2$	$(\bar{X}_i - \bar{X}_G)^2$
Group 1	-2	-1	0	4	1	1
	-1	-1	0	1	0	1
	0	-1	0	0	1	1
Group 2	-1	0	0	1	1	0
	0	0	0	0	0	0
	1	0	0	1	1	0
Group 3	0	1	0	0	1	1
	1	1	0	1	0	1
	2	1	0	4	1	1
			12	6	6	
			$SS_T$	$SS_W$	$SS_B$	

Table 15.2: Sums of squares for the hypothetical data from Table 15.1.

$$SS_T = SS_W + SS_B$$

which is the fundamental equation of ANOVA—the unique partitioning of the total sum of squares ( $SS_T$ ) into two components: the sum of squares within groups ( $SS_W$ ) plus the sum of squares between groups ( $SS_B$ ).

These computations for the earlier data are shown in Table 15.2. Note that for  $SS_B$ , each group mean is used in the calculation three times—once for each score that its corresponding group mean contributes to the total. In general, then, using  $n_i$  as the number of scores in each group,  $SS_B$  also may be written as:

$$SS_B = \sum_i n_i (\bar{X}_i - \bar{X}_G)^2$$

### 15.1.2 Mean squares

Sums of squares are not much use in and of themselves because their magnitudes depend *inter alia* on the number of scores included in producing them. As a consequence, we normally use the *mean squared deviation* or *variance* (see Chapter 3). Because we are interested in estimating population variances, these *mean squares* ( $MS$ ) are calculated by dividing each sum of squares by the *degrees of freedom* ( $df$ )<sup>6</sup> associated with it. For the mean square between groups ( $MS_B$ ), the associated degrees of freedom is simply the number of groups minus one. Using  $a$  to represent the number of groups:

$$df_B = a - 1$$

<sup>6</sup>Degrees of freedom, as you no doubt recall, represent the number of values entering into the sum that are free to vary; for example, for a sum based on 3 numbers, there are 2 degrees of freedom because from knowledge of the sum and any 2 of the numbers, the last number is determined.

For the mean square within groups ( $MS_W$ ), each group has  $n_i - 1$  scores free to vary. Hence, in general,

$$df_w = \sum_i (n_i - 1)$$

and with equal  $n$  per group:

$$df_W = a(n - 1)$$

Note that:

$$df_T = df_B + df_W$$

Consequently, (for equal  $n$ ),

$$MS_W = \frac{SS_W}{a(n - 1)}$$

and

$$MS_B = \frac{SS_B}{(a - 1)}$$

$MS_W$  is simply the pooled estimated variance of the scores, which, according to the null hypothesis, is the estimated population variance. That is,

$$MS_W = \hat{\sigma}_x^2$$

To estimate the variance of the *means* of samples of size  $n$  sampled from the same population, we could use the standard formula:<sup>7</sup>

$$\hat{\sigma}_x^2 = \frac{\hat{\sigma}_x^2}{n}$$

which implies that:

$$\hat{\sigma}_x^2 = n\hat{\sigma}_x^2$$

But,

$$MS_B = n\hat{\sigma}_x^2$$

Hence, given that the null hypothesis is true,  $MS_B$  *also* is an estimate of the population variance of the scores.

---

<sup>7</sup>This expression is similar to the formula you encountered in Chapter 12, except that it is expressed in terms of *variance* instead of *standard deviation*; hence, it is the *estimated variance of the mean* rather than its *standard error*.

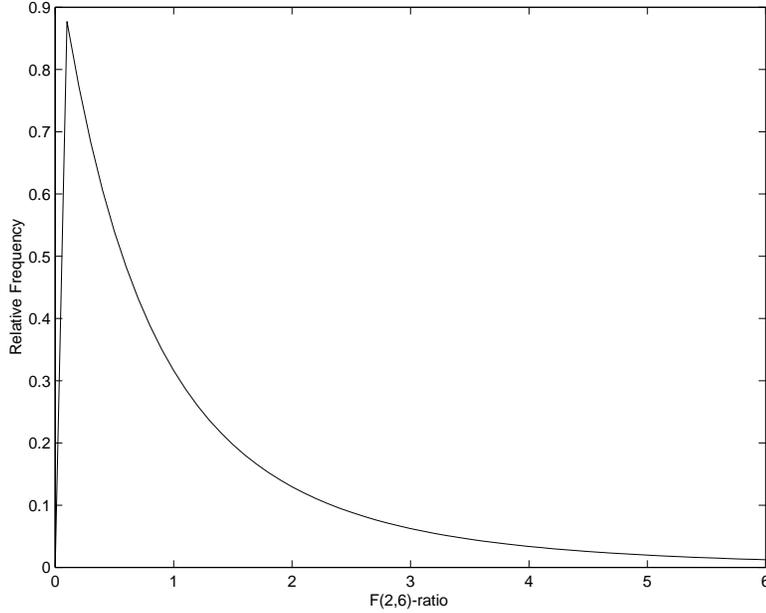


Figure 15.1: Plot of the  $F$ -distribution with 2 and 6 degrees of freedom.

### 15.1.3 The $F$ -ratio

We define the following  $F$ -ratio:<sup>8</sup>

$$F(df_B, df_W) = \frac{MS_B}{MS_W}$$

If the null hypothesis is true, then this ratio should be approximately 1.0. Only approximately, however; first, because both values are estimates, their ratio will only tend toward their expected value; second, even the expected value—long-term average—of this ratio is unlikely to equal 1.0 both because, in general, the expected value of a ratio of random values does not equal the ratio of their expected values, and because the precise expected value under the null hypothesis is  $df_W / (df_W - 2)$ , which for the current example is  $6/4 = 1.5$ .

Of course, even if the null hypothesis were true, this ratio may deviate from its expected value (of approximately 1.0) due to random error in one or the other (or both) estimates. But the *probability distribution* of these ratios can be calculated (given the null hypothesis assumptions mentioned earlier). These calculations produce a family of distributions, one for each combination of numerator and denominator degrees of freedom, called the  $F$ -distributions. Critical values of some of these distributions are shown in Table 15.3.

<sup>8</sup>Named in honour of Sir Ronald A. Fisher. Fisher apparently was not comfortable with this honorific, preferring to call the ratio  $Z$  in his writing.

The  $F$ -distributions are positively skewed, and become more peaked around their expected value as either the numerator or the denominator (or both) degrees of freedom increase.<sup>9</sup> A plot of the probability density function of the  $F$ -distribution with 2 and 6 degrees of freedom is shown in Figure 15.1. If the null hypothesis (about equality of population means) isn't true, then the numerator ( $MS_B$ ) of the  $F$ -ratio contains variance due to differences among the population means in addition to the population (error) variance resulting, on average, in an  $F$ -ratio greater than 1.0. Thus, comparing the obtained  $F$ -ratio to the values expected theoretically allows one to determine the probability of obtaining such a value if the null hypothesis is true. A sufficiently improbable value (i.e., much greater than 1.0) leads to the conclusion that at least one of the population means from which the groups were sampled is different from the others.

Returning to our example data,  $MS_W = 6/(3(3 - 1)) = 1$  and  $MS_B = 6/(3 - 1) = 3$ . Hence,  $F(2, 6) = 3/1 = 3$ . Consulting the appropriate  $F$ -distribution with 2 and 6 degrees of freedom (see Figure 15.1 and Table 15.3), the critical value for  $\alpha = .05$  is  $F = 5.14$ .<sup>10</sup> Because our value is less than the critical value, we conclude that the obtained  $F$ -ratio is not sufficiently improbable to reject the null hypothesis of no mean differences.

## 15.2 $F$ and $t$

Despite appearances, the  $F$ -ratio is not really an entirely new statistic. Its theoretical distribution has the same assumptions as Student's  $t$ -test, and, in fact, when applied to the same situation (i.e., the difference between the means of two independently sampled groups),  $F(1, df) = t_{df}^2$ —that is, it is simply the square of  $t$ .

## 15.3 Questions

1. You read that “an ANOVA on equally-sized, independent groups revealed an  $F(3, 12) = 4.0$ .” If the sum-of-squares between-groups was 840,
  - (a) Complete the ANOVA summary table.
  - (b) How many observations and groups were used in the analysis?
  - (c) What should you conclude about the groups?
2. While reading a research report in the *Journal of Experimental Psychology* (i.e., some light reading while waiting for a bus), you encounter the following statement: “A one-way ANOVA on equally-sized, independent groups revealed no significant effect of the treatment factor [ $F(2, 9) = 4.00$ ]”.

<sup>9</sup>Because the accuracy of the estimate increases as a function of  $n$ .

<sup>10</sup>That is, according to the null hypothesis, we would expect to see  $F$ -ratios as large or larger than 5.14 less than 5% of the time.

$df_2$	$df_1$									
	1		2		3		4		5	
	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
1	161.45	4052.2	199.5	4999.5	215.71	5403.40	224.58	5624.60	230.16	5763.60
2	18.51	98.50	19.00	99.00	19.16	99.17	19.25	99.25	19.30	99.30
3	10.13	34.12	9.55	30.82	9.28	29.46	9.12	28.71	9.01	28.24
4	7.71	21.20	6.94	18.00	6.59	16.69	6.39	15.98	6.26	15.52
5	6.61	16.26	5.79	13.27	5.41	12.06	5.19	11.39	5.05	10.97
6	5.99	13.75	5.14	10.93	4.76	9.78	4.53	9.15	4.39	8.75
7	5.59	12.25	4.74	9.55	4.35	8.45	4.12	7.85	3.97	7.46
8	5.32	11.26	4.46	8.65	4.07	7.59	3.84	7.01	3.69	6.63
9	5.12	10.56	4.26	8.02	3.86	6.99	3.63	6.42	3.48	6.06
10	4.96	10.04	4.10	7.56	3.71	6.55	3.48	5.99	3.33	5.64
15	4.54	8.68	3.68	6.36	3.29	5.42	3.06	4.89	2.90	4.56
20	4.35	8.10	3.49	5.85	3.10	4.94	2.87	4.43	2.71	4.10
30	4.17	7.56	3.32	5.39	2.92	4.51	2.69	4.02	2.53	3.70
40	4.08	7.31	3.23	5.18	2.84	4.31	2.61	3.83	2.45	3.51
50	4.03	7.17	3.18	5.06	2.79	4.20	2.56	3.72	2.40	3.41

Table 15.3: The critical values of the  $F$ -distribution at both  $\alpha = 0.05$  and  $\alpha = 0.01$  as a function of numerator ( $df_1$ ) and denominator ( $df_2$ ) degrees of freedom.

- (a) How many groups were there in the analysis?
- (b) How many observations were used in the analysis?
- (c) Given that the mean-square within groups was equal to 90, complete the ANOVA summary table.
- (d) In more formal language, what is meant by the phrase “no significant effect of the treatment factor”?

Part III

The Psychological  
Literature



## Chapter 16

# The Literature

There are several formal ways of making the results of scientific research public. They include presentations at professional meetings, journal articles, and books. Each serves a slightly different purpose, and the procedure is somewhat different but in all cases the main purpose is identical: Sharing the results of one's research with others.<sup>1</sup>

If you don't make your research results public, you aren't doing science—you may be doing research, research and development, playing mental games, or something else, but one of the main criteria of science is that the findings are public.

One of the main goals of this course is to help you gain the skills necessary to comfortably read the original scientific psychology literature. For the most part, this means articles in psychology journals, although when you are first entering an area, or looking for an overview, books may be useful and for the most up-to-date information scientific meetings are preferred.

### 16.1 Meetings

Academics attend professional meetings in order to tell others in their field what they are doing and to find out what the others are doing. These activities are an important part of the research process and of the informal lines of communication that build up around any area of expertise. Professional meetings give you the most up-to-date information possible if you don't know someone working in the area well enough to phone them up and say "so what's new?".

---

<sup>1</sup>That is, in addition to just telling one's friends, which is an important source of current information for those who are "in the club". We are, after all, talking about a relatively small group once you get to a reasonable degree of specialisation.

Examples of yearly conferences that serve this purpose are the annual meetings of: the American Psychological Association (APA), the Society for Experimental Social Psychology (SESP), the Canadian Psychological Association (CPA—important for social psychologists and clinical psychologists), the Psychonomic Society (important for cognitive psychologists), the Canadian Society for Brain, Behaviour, and Cognitive Science. (CSBBCS or BBCS<sup>2</sup>—important for cognitive psychologists and neuroscientists), the Society for Neuroscience, and Advances in Research in Vision and Ophthalmology (ARVO). There are also regional conferences (e.g., Banff Annual Seminar in Cognitive Science—BASICS) and smaller, more specialized conferences (e.g., Society for Computers in Psychology—SCiP, Consciousness and Cognition).

An important reason for keeping up on the latest research by attending conferences is so that you don't duplicate the work of other researchers—you don't want to have two labs working on the same project. There will, of course, be a lot of overlap between the research you choose to do and the research that is taking place in other labs but life is too short to waste your time doing what someone else is doing. You want your research to be informed by the latest related findings. You don't want to be working on a project based on a certain assumption only to find out that another researcher has just proven that assumption wrong.

Much of the benefit of conferences comes from the informal discussions one has with colleagues over dinner or drinks, in shared cabs, while waiting in airports, and so on. Many of the best insights and ideas that emerge from conferences come from these intense, one-on-one discussions. It is this availability of colleagues in close proximity and in the absence other obligations (e.g., teaching, committee meetings, family obligations) that ensures the survival and effectiveness of these conferences in the days of email and virtual conferencing. In addition to the critical informal discussions, there are two formal types of presentations that generally occur and often serve as the springboard for later, informal discussions. They are talks and posters.

### 16.1.1 Talks

Research talks generally take the form of a short lecture laying out the results of a research project. They allow researchers efficiently to disseminate their most recent work to relatively large groups of interested colleagues. The presenter, usually the primary researcher on the project, lays out the reason for undertaking the project, how the project was carried out, what was found, and what it means in terms of psychological theories. As the time is usually limited (never long enough in the opinion of the presenter, although occasionally far too long in the opinion of the listeners), the information is presented in substantially less detail than would be contained in a journal article on the same research. This lack of detail is not a problem because the listeners are generally given the opportunity

---

<sup>2</sup>Leave it to a group of Canadian scientists to give themselves a name so long that they even have to shorten the acronym.

to ask questions about any additional information they desire. There is always a short time scheduled for questions at the end of the talk, and the presenter is usually available at the conference for several days.

### 16.1.2 Posters

Posters are visual displays (generally about 4ft X 8ft in size) that present much the same information as is presented in a talk. They allow a lot of people to present their research in a short period of time. Poster sessions generally run for about two hours in rooms containing anywhere from a few dozen to several hundred posters (depending on the size of the conference). In that time the presenter hangs around in front of his or her poster while people mill about, read the poster (or pass it by), and discuss it with the presenter. The poster format allows for more two-way interaction between the presenter and his or her audience than a talk does and for greater flexibility in how the information is presented. The information can be tailored to the individual to whom it is being presented—the researcher can discuss esoteric details of procedure and theory with other experts in the area and general issues and relations to other areas of psychology with experts from other areas. The best posters use lots of visual information—pictures of apparatus and procedures, and graphs of data—and use little writing. Such posters allow the reader quickly to get the gist of research and then decide whether to follow up on details with a closer look or by talking to the presenter. Although the audience for a single poster is often smaller than for a single talk, the interaction with each member of the audience is generally more intense for posters.

Some research is more easily presented as a poster, and some as a talk. Either way, though, the important thing is that the research is out in the open being put to the test of public scrutiny where other researchers can criticize the methods or assumptions being made, argue with the conclusions being drawn, and generally help to keep the researcher on the intellectual “straight and narrow”. In addition, good research will influence the work of others in the field, hastening the advance of the science as a whole. Remember, science is a public process.

## 16.2 Journals

The main outlet for information in any field of science is in the field’s journals. Journals are regularly published (often bimonthly or quarterly) magazines containing detailed reports of original research, and occasionally, review articles that summarize the state of some specific field of research. There are hundreds of psychology journals, including dozens of high-quality, respected ones. There are journals of general interest to many scientific psychologists, journals devoted to specific areas, journals devoted to even more specific areas, and journals devoted to reviews of various areas of psychology.

All journals are edited by an expert in the field. It is the editor’s job to ensure that only high-quality papers of interest to the readers of his or her

journal get published in it. Selecting the appropriate papers is accomplished by sending the submitted manuscripts to experts in the field for their comments and criticisms before deciding whether or not to publish those manuscripts—a process known as peer review. If the other experts find the paper sufficiently interesting, sufficiently well-written, and free of flaws of logic or experimental design, the paper will likely be published. If the paper is deficient in one or more of these areas in a way that the editor feels can be remedied, it will likely be sent back to the author(s) for revision. Otherwise, the paper will be rejected and the readers of the journal will have been protected from having their time and money wasted reading poor quality research.

### 16.3 Books

Books tend to be more general than journal articles, covering a topic more broadly. Because of the time required to write a book, or to coordinate several different authors writing individual chapters in the case of edited books, books are usually less up-to-date than journal articles. They are often a good starting point when entering a new area because they afford the ability to cover a broad area with greater detail than there is room for in a journal article. Keep this in mind when you are researching an area. Any research report or student project based entirely on books will be inadequate due to the dated nature of the information available in books.

## Chapter 17

# The APA Format

Virtually all of the journal articles you will read (or write) will be in APA format. Historically, many of the leading journals in areas of scientific psychology were published by the APA, and it is the APA, consequently, that controls the rules of writing for those journals. APA is the TLA<sup>1</sup> for the American Psychological Association, the association that, at one time, was meant to represent all psychologists. More recently, most of the scientific psychologists have left the APA to form the American Psychological Society (APS), leaving the APA mainly to the clinical psychologists. The Canadian equivalents of these two organisations are the Canadian Psychological Association (CPA—for clinicians) and the Canadian Society for Brain, Behaviour, and Cognitive Science (CSBBCS—for scientists). The British equivalents are the British Psychological Association (BPA—for clinicians) and the Experimental Psychology Society (EPS—for scientists). Many journals that are not published by the APA have also adopted the APA guidelines or slightly modified versions of them. These journals include virtually all psychology journals and many others in related fields. Journals in neuroscience are more likely to use other forms of referencing, but you'll note that the structure of the articles is similar to APA format.

APA format is a set of rules for writing journal articles that are designed to make the articles easy for the publisher to publish and easy for the reader to understand. A brief overview of these rules is given here. The specifications for writing in APA format—in their full, detailed, glory—can be found in the *Publication Manual of the American Psychological Association, Fifth Edition* (American Psychological Association, 2001). When reading journal articles it will be very helpful to know the format in which they are written. When writing journal articles or class assignments based on them (i.e., research reports, research proposals) it is imperative that you know the format.

Specifications for the manuscript itself (e.g., line spacing, margin sizes, headers on the manuscript, placing of tables, figures, and footnotes, etc.) are largely for the benefit of those who work in the publishing process: Double spacing

---

<sup>1</sup>Three Letter Acronym

and large margins allow for easy placing of copy editor's comments; consistent line spacing and margins make it easy to estimate how many journal pages a manuscript will take; headers and page numbers on a manuscript make it easy to keep track of the pages; and separate tables, figures, and footnotes are ready to be easily copied, and set in type.

Specifications for the structure of the article (e.g., order of the sections, content of the sections) are largely for the benefit of the reader. Anyone reading a journal article in an APA journal can be sure that certain types of information will be in certain portions of the article. This consistency makes it much easier to quickly skim articles or to scan an article looking for specific information. The different sections of an APA formatted research article and their contents are described in the following sections.

### 17.0.1 Title

The title is probably the most important set of words in your article and the one least considered by most students (you know you're not going to get a good mark on a research proposal that you've titled *Research Proposal*). It is the title that is likely to attract people to read your article<sup>2</sup> and so most researchers spend weeks, if not months considering just the right title for their article. The ideal title will be short, informative, and clever. Although all three of these criteria may not be compatible in every case, you should strive for at least two of three. One of my favourites is an article by Ian Begg (1987) about the meaning ascribed to a particular logical term by students entitled *Some*.

### 17.0.2 Abstract

The abstract is the second most important set of words in your article. It is a short summary of your entire article. Abstracts typically run from 75 to 150 words. The specific maximum length for the abstract is set by the policy of the particular journal in which the article appears. The abstract is critical because it is the part of the article that is most likely to be read. In fact, it is the only portion of the article that will be read by the vast majority of people who are intrigued enough by the title to read further. The abstract is also, at least at present, easier to obtain than any other part of an article. Most online databases contain the abstracts of all the articles they list, but not the full articles. Thus, great care should be taken in writing the abstract to ensure that it is as informative as possible within the constraints of the specified maximum word count. The abstract should state what theories were tested, what experimental manipulations were made, and what the results of those manipulations were.

---

<sup>2</sup>unless you've built such a reputation as a researcher and/or writer that people will read the article on the basis of your name alone—good luck.

### 17.0.3 Introduction

The introduction is pretty much what it sounds like. It is in the introduction that the author introduces the problem that he or she is trying to solve. An introduction generally starts out by reviewing the relevant literature. That is, summarizing what other researchers have discovered about the general topic of the article. Then, the author's particular point of view is introduced along with supportive evidence from the relevant literature. The specific experimental manipulation that was carried out is introduced along with how it will add to the collective knowledge about the topic at hand. Most of the time an introduction will start off relatively general and become more and more specific as it goes along until it narrows in on the precise manipulations that will be carried out and their potential meaning to the relevant literature.

It is helpful when reading a research article—and even more helpful when writing one—to think of the article as a story. The writer's job is to tell a coherent, convincing story based on the research findings in the field in which the research takes place, and on the new research findings presented in the article. Thus, all the writing in the article is part of the story and any information that doesn't help to move the story along shouldn't be there. Try to keep this point in mind when you are constructing the literature review sections of any papers you write. It will help you avoid the common error of simply listing a lot of research findings without explaining the connection of each to the purpose of your paper (i.e., what role they play in the story you are trying to tell).

### 17.0.4 Method

The method is where the specific methods used in the study are explained. A reader of the article should be able to replicate the study (i.e., do it again in exactly the same way) using only the information in the method section. If that is not possible (for example, if you had to contact the author to ascertain some additional details) the method section was not properly written. The method section is generally broken down into subsections. The specific subsections that appear will vary with the type of the experiment, but the following subsections are generally used.

#### **Participants**

The participants subsection contains particulars about who took part in the study. Who participated, how many were there, how were they recruited, and were they remunerated in some way? If animals were used, what kind were they, where were they obtained, and did they have any specific important characteristics? For most studies that use “normal” people as participants this section will be quite short. However, if special populations are used, for example, people with particular types of brain damage or animals with unusual genetic structures, the method section can be quite long and detailed.

### **Materials**

The materials section contains descriptions of any important materials that were used in the study. Materials can be quite varied including things like lists of words, stories, questionnaires, computer programs, and various types of equipment. In general, anything physical that is used in the study is described in the materials section (including things such as answer sheets and response buttons). Anything that is purely mental (e.g., counting backward by threes, imagining an elephant) is described in the procedure section.

### **Procedure**

The procedure section contains precisely what was done in the study: What steps were taken, and in what order. The best procedure sections tend to take the reader through the study step by step from the perspective of a participant. It is important that the reader have a good feeling for precisely what was done in the study, not only for the sake of replicability but in order to understand what a participant is faced with and consequently might do.

### **Details, details**

Remember, when deciding what to include in the method section, that the overriding principle is to give enough information in enough detail that a reader of the paper could replicate the study—and no more. Thus, including “HB pencils” in the materials section is probably not important (unless the study deals with the perceptibility of handwriting), likewise listing 23 females with a mean age of 18.6 years and 17 males with a mean age of 19.2 years is also probably overkill (unless the study looks at sex differences on the task as a function of age). On the other hand, if one half of your participants were recruited at a local high school and the other half at a senior citizen’s centre, you should probably mention it.

Don’t forget to include in your method section what you are measuring and what you will be analyzing in the results section.

## **17.0.5 Results**

The results section contains what was found in the study (what it means goes in the discussion section, the reason for looking for it goes in the introduction). If you have a lot of numbers to deal with, a figure or table is often the best way to present your data. It is important to remember that you are trying to tell a story and that the data are part of the story. Good story-telling means that you shouldn’t just present a whole bunch of numbers and let the reader figure it out. You should present the numbers that are important to your story, and where a difference between sets of numbers is important, you should statistically test that difference and report the results of the test. The statistical tests are usually reported in the body of the text rather than on the graphs or in the the tables.

You want to start out by stating the type of statistical test you are using and the alpha  $p$  value you are using (usually  $p < .05$ ). Don't forget to tell the reader what you were measuring (what the numbers stand for) and what the direction of any differences was. You want to report the results of any statistical tests as clearly and concisely as possible. You can usually get a good idea of how to report the results of statistical tests by looking at how the authors in the papers you've been reading do so. Here are a couple of examples of how to report the results of a statistical test:

... There were significantly higher testosterone levels in the males (68.5 ppm) than in the females (32.2 ppm),  $t_{(34)} = 7.54$ . ...

... The final weight of the participants was subjected to a 2 x 2 between subjects analysis of variance (ANOVA) with sex of the subjects and age of the subjects as factors. Males (mean = 160 lb) weighed significantly more than females (mean = 140 lb),  $F_{(1,35)} = 5.32$ ,  $MSe = 5.5$ . Older subjects (mean = 148 lb) did not weigh significantly more than younger subjects (mean = 152 lb),  $F_{(1,35)} = 1.15$ ,  $MSe = 5.5$ , n.s. Finally, there was no significant interaction between age and sex of the subjects ( $F < 1$ ). ...

Note that the direction of any differences is noted, as are the statistical significance of the differences, the values of the statistics, and means for each group. In these examples there is no need for any tables or figures as the information is all presented in the body of the text. Don't spend a lot of time explaining the meaning of the results in this section; that comes in the discussion section.

## 17.0.6 Discussion

The last section of the body of the paper is where you explain what your results mean in terms of the questions you set out to answer—where you finish the story. The first part of the discussion generally summarizes the findings and what they mean and the later parts generally deal with any complications, any new questions raised by the findings, and suggestions for future directions to be explored.

## 17.0.7 References

The reference section is the most complex part to learn how to write but often the most informative part to read—it leads you to other related articles or books. The reference section lists all the sources you have cited in the body of the paper and it lists them in a complete and compact form. Every source of information cited in the body of the paper must appear in the reference section, and every source in the reference section must be cited in the body of the paper.

There are specific ways of referencing just about any source of information you could imagine. For most, however, it is easy to think of a reference as a paragraph containing three sentences. The three most common sources of information are journal articles, chapters in edited books, and entire books written by a single author or group of authors. In all these cases the reference

consists of three sentences, the first describes the author(s), the second gives the title of the work, and the third tells you how to find it. In the case of a journal article the third section contains the title of the journal and volume and issue in which it is contained. In the case of a chapter in an edited volume, the third section contains the name of the book, the name of the editor, and the publisher. In the case of a book, the third section contains the publisher of the book. Examples of the three types of references follow:

References in APA format.

Journal Article:

Vokey, J. R., & Read, J. D. (1985). Subliminal messages: Between the Devil and the media. *American Psychologist*, *40*, 1231–1239.

Chapter in an edited book:

Allen, S. W. & Vokey, J. R. (1998). Directed forgetting and rehearsal on direct and indirect memory tests. In J. Golding & C. MacLeod (Eds.) *Intentional forgetting: Interdisciplinary approaches*. Hillsdale NJ: Erlbaum.

Book:

Strunk, W., Jr., & White, E. B. (1979). *The elements of style (3rd ed.)*. New York: Macmillan.

Note in particular the punctuation and formatting of the example references. Each of the three sentences ends with a period. The titles of journal articles, book chapters, and books are capitalized only in the way that a regular sentence would be capitalized. Only the titles of journals have all the main words capitalized.

Examples of citations in the text to the references listed above.

According to Vokey and Read (1985) . . .

. . . as shown by the very clever (and good-looking) Allen and Vokey (1998).

Of course, one should almost always use “that” rather than “which” (Strunk & White, 1979). That (not which) is why several style experts suggest going on a “which” hunt when proof reading.

### 17.0.8 Notes

When writing the paper all notes, including footnotes, and author notes (a special type of footnote that gives contact information for the author and acknowledges those who have helped with the paper) appear at the end of the manuscript

immediately following the reference section. Footnotes should be used sparingly and should contain information (often details) that is important but would, if contained in the body of the text, interrupt the flow of the story.

### **17.0.9 Tables**

Any section of text or numbers that is set apart from the main body of the text is called a table. Tables are most commonly used to present data (sets of numbers) but are occasionally used to display examples of materials. Each table requires an explanatory caption and must be referred to and described in the body of the article. Each table is typed, along with its caption, on a separate page. The tables appear, in the order in which they are referenced in the body of the article, immediately following the footnotes page.

### **17.0.10 Figures**

A figure is information of a pictorial nature. The most common use of figures is to display data in a graph, although they are less commonly used to show drawings or pictures of apparatus or to summarize procedures. As with tables, figures are consecutively numbered in the order in which they are referenced in the body of the article. Also like tables, each figure requires an explanatory caption and must be referred to and described in the body of the article. However, unlike tables, the captions for the figures are printed separately from the figures on the page immediately following the tables but immediately preceding the figures themselves.

### **17.0.11 Appendices**

An appendix contains something that is important but that is peripheral enough to the main point of the article that it can be put at the end in a separate section. The most common uses of appendices are for sets of materials (e.g., some interesting set of sentences with weird characteristics, or a questionnaire) or complex statistical points that would break up the flow of the paper if left in the main part (e.g., the mathematical derivation of a new statistic).



**Part IV**  
**Appendices**



# Appendix A

## Summation

Most of the operations on and summaries of the data in this book are nothing more than various kinds of sums: sums of the data themselves and sums of different transformations of the data. Often these sums are *normalised* or corrected for the number of items summed over, so that the now-normalised sum directly reflects the magnitude of the items summed over, rather than being in addition confounded with the sheer number of items making up the sum. But even then, these normalised sums are still essentially sums. Because sums are central to the discussion, it is important to understand the few, simple properties of sums and the related principles of summation.

### A.1 What is Summation?

Summation is the adding together of numerical items in a specified set of numerical items. We specify the set to be summed by a name, typically a single letter. Subscripted versions of the letter then index the specific items or numbers in the set. For example, for the set  $X = \{8, 4, 8, 2\}$ , the name of the set is  $X$ ,  $X_1$  refers to the first member of the set, or 8,  $X_4 = 2$ , and, in general,  $X_i$  refers to the  $i$ th member of the set. If the set  $X = \{1, 2, 3, 4, 5\}$ , then the sum of  $X$ —that is, the sum of all the items within  $X$ —is equal to  $15 = 1 + 2 + 3 + 4 + 5$ .

#### A.1.1 The Summation Function

In many computer spreadsheets, summation can be accomplished by concatenating the cell references with “+” signs in the formula for the cell that is to hold the sum, as in, for example, `=B1+B2+B3+B4` to compute the sum of the values in cells B1 through B4 of the spreadsheet. However, it is almost always also provided as *function* to return a value (the sum) from an input (the items of the set) that is a specified function (in this case, summation) of the input. Where summation is provided as a built-in rather than a programmed function, it often

is written as the cell formula =SUM(B1:B4), indicating that the sum of the items in cells B1 through B4 is to be the *value* of the cell containing the formula.

### A.1.2 The Summation Operator: $\Sigma$

Mathematicians, of course, can't be denied their own stab at the arcane (indeed, their nomenclature or terminology often exemplifies the word). In the original Greek alphabet (i.e., alpha, beta, ...), the uppercase letter corresponding to the Arabic (or modern) alphabet letter "S"—for "sum"—is called sigma, and is written as:  $\Sigma$ . To write "the sum of  $X$ " in mathematese, then, one writes:

$$\Sigma X$$

which means "sum all the items of  $X$ ".

More formally, and to allow for the possibility that one might want to sum over some limited range of items of  $X$  (say, the second to the fourth), the sum is often designated with an explicit index to denote which items of  $X$  are to be summed over. To designate that the sum is over only the second to fourth items, for example, one would write:

$$\sum_{i=2}^4 X_i$$

Although it is redundant (and, when did that ever hurt?), the sum over all items, where  $n$  denotes the number of items of  $X$ , is often written as:

$$\sum_{i=1}^n X_i$$

An individual item of  $X$  is designated by its *index*, or *ordinal position* within  $X$ . Note that by ordinal position, we do not mean ordinal *value*, or *rank*. In the set  $X = \{1, 2, 3, 4, 5\}$  ordinal position and ordinal value happen to be equivalent (e.g., the second item also happens to have the value of 2, which is also the value of the second item when ordered by rank; that is, value, ordinal position, and ordinal value are all the same). In the set  $Y = \{3, 5, 1, 6\}$ , however, the second item has a value of 5, which is neither its ordinal position (2) or its ordinal value

(3). If  $X = \{6, 9, 1, 2, 7, 7\}$ , then,  $X_4 = 2$ ,  $n = 6$ , and  $\sum_{i=1}^6 X = 6 + 9 + 1 + 2 + 7 + 7$ .

## A.2 Summation Properties

All of the foregoing provides us with a convenient method of designating sums and the items going into them, but does little to explicate the properties and principles. As sums are just the results of addition, they embody all of the properties of addition.

### A.2.1 Summation is commutative

Probably the most useful of these properties is the fact that addition and, thereby, summation is commutative. To be commutative is to have the property that the order of the operation does not affect the result. No matter what order we sum the items of  $X = \{1, 2, 3, 4, 5\}$ , the result is always equal to 15. Multiplication (which may be characterised as addition done quickly, at least for integer operands: e.g.,  $4 * 5 = 5 * 4 = 5 + 5 + 5 + 5 = 4 + 4 + 4 + 4 + 4 = 20$ ) is also commutative. Subtraction is not, nor is division (which, analogously, is just subtraction done quickly). However, as subtraction can be rendered as the addition of negative numbers [e.g.,  $4 - 5 = 4 + (-5) = -5 + 4 = -1$ ], the noncommutativity of subtraction is easily avoided. Similarly, division can be rendered as multiplication with negative exponents (e.g.,  $4/5 = 4 * 5^{-1} = 5^{-1} * 4 = 0.8$ ). Consequently, the noncommutativity of division can also be avoided. In this sense, all arithmetic operations reduce to addition, and share in its properties. Indeed, from this perspective, subtraction, multiplication, and division (and also exponentiation) can be seen as no more than shorthand, hierarchical, re-writing rules for what would otherwise be tedious written statements of explicit addition (at least for integer operands: for fractions and especially irrational real numbers, these simple parallels no longer necessarily hold).<sup>1</sup> At any rate, once rendered as a sum, including using the negative number and negative exponent notation for subtraction and division, the entire operation is commutative or order independent.<sup>2</sup>

### A.2.2 Summation is associative

The associative property of addition is what gives multiplication (as rapid addition or re-writing rule) its force. Both addition and multiplication are associative (and, as noted earlier, both subtraction and division can be if written appropriately). This property is derived from the property of being commutative, but is often discussed separately because it emphasizes something not immediately obvious from the commutative property, that of *partial sums* or *partial products*. Any sum can be represented as the sum of partial sums. For example, with parentheses constraining the order of addition for the numbers 1, 2 and 3,  $(1 + 2) + 3 = 3 + 3 = 1 + (2 + 3) = 1 + 5 = 6$ . Any sum, then, can be partitioned into the sum of partial sums, an important property that we will exploit in detail subsequently.

For the moment, though, we will note that it is this property that is exploited

---

<sup>1</sup>Exponentiation—raising numbers to powers—may similarly be characterised as a rewriting rule for multiplication. Viewing multiplication as just a rewriting rule for addition—as is the case for everyday encounters with addition—implies a limited or reduced form of arithmetic known as *Presburger arithmetic*. A more complex form of arithmetic—known as *Peano arithmetic*—includes multiplication as a distinct concept independent of addition. For our current purposes, Presburger arithmetic is all that is needed. For those with an interest in esoterica, it is Peano arithmetic that is assumed as a premise in Gödel’s famous incompleteness theorem; Presburger arithmetic is complete in this mathematical sense.

<sup>2</sup>A related property is that addition is *distributive*:  $a - (b + c) = a - b - c$

in most computer (and “by-hand”) algorithms for computing sums, including the one most people use. However, there is another way of thinking about this property. To add or sum the numbers of  $X$ , we usually take each number of  $X$  in succession and add it to the accumulated sum of the numbers preceding it. That is to say, the sum of a set of numbers can be defined recursively—in terms of itself:

$$\sum_{i=1}^n X_i = \sum_{i=1}^{n-1} X_i + X_n$$

In English, the sum of the numbers from the first to the last is equal to the sum of the numbers from the first to last but one (the “penultimate” number— isn’t wonderful that we have words for such abstruse concepts?) plus the last number. But what is the sum of the numbers from the first to the penultimate number? It is the sum of the numbers from the first to the penultimate but one, plus the penultimate number, and so on, until we reach the the sum of the first number, which is just the first number.

### The secret to living to 100 years of age

Note how this summation function recursively calls itself (i.e., it uses itself in its own definition). If you find this idea of recursion somewhat bewildering, consider the question of how to live to be 100 years of age. Clearly, the answer is to live to be 99, and then to be careful for one year. Ah, but how does one live to be 99? Clearly, live to be 98 and then be careful for one year, and so on.<sup>3</sup> The associative property of summation is like that: summation is just the summing of sums. Ah, but how do you get the sums to summate? Clearly, by summation!

## A.3 Summation Rules

It is the associative property that allows for the simple addition of pre-existing sums to determine the sum of the combination of their respective items. It is also the associative property that leads to the first summation rule.

### A.3.1 Rule 1: $\sum(X + Y) = \sum X + \sum Y$

This rule should be intuitively obvious; it states that the sum of the numbers over the two sets  $X$  and  $Y$  is equal to the sum of the sums of  $X$  and  $Y$ , and follows from the fact that, for example, the sum of  $\{1, 2, 3\} + \{4, 5, 6\} = (1 + 2 + 3) + (4 + 5 + 6) = 6 + 15 = 1 + 2 + 3 + 4 + 5 + 6 = 21$ . However, it is a handy re-writing rule to eliminate the parentheses in the expression.

---

<sup>3</sup>One could also ask how to be careful for one year, and the answer is just as clearly to live for 364 days, and then be careful for one day, and so on. And how does be careful for one day? Live for 23 hours and then be careful for one hour . . . .

**A.3.2 Rule 2:**  $\sum(X - Y) = \sum X - \sum Y$ 

Similar to rule 1, it makes no difference whether you subtract first and then sum  $[\sum(X - Y)]$  or sum first and then subtract the respective sums  $(\sum X - \sum Y)$ , as long as the ordering is maintained. If, as discussed in A.2.1, subtraction is re-written as the sum of negative numbers, then rule 2 becomes rule 1.

**A.3.3 Rule 3:**  $\sum(X + c) = \sum X + nc$ 

The associative property also allows for the simple computation of the new sum following the addition of a constant to every item of  $X$ . Suppose, for example, it was decided after set  $X$  (consisting, say, of hundreds or thousands of items) had been laboriously summed that a value of  $c$  should be added to every score (where  $c$  could be either negative or positive). One approach would be to ignore the computed sum of  $X$ , and to add  $c$  to every item of  $X$ , producing a new set of scores  $Z$ , and then sum the items of  $Z$ . Workable, but tedious.

The associative property, however, suggests another, more efficient approach. For every item  $i$  of  $X_i$ ,

$$Z_i = X_i + c$$

So,

$$\sum_{i=1}^n Z_i = (X_1 + c) + (X_2 + c) + \cdots + (X_n + c)$$

But, by the associative property, this sum could be re-written as:

$$\sum_{i=1}^n Z_i = \underbrace{(X_1 + X_2 + \cdots + X_n)}_{\sum X} + \underbrace{(c + c + \cdots + c)}_{nc}$$

As there is one  $c$  for each item in  $X$ , there are a total of  $n$  values of  $c$  being added to the original sum:

$$\sum_{i=1}^n Z_i = \sum_{i=1}^n X_i + nc$$

which is just the original sum plus  $nc$ . So, if you have the sum for, say, 200 scores, and wish to compute the sum for the same scores with 5 added to each one (e.g., some generous stats prof decided to give each of 200 students 5 bonus marks), the new sum is simply the old sum plus  $200 * 5 =$  the old sum +1000. Cool. The whole operation can be summed up (pun intended!) in the following expression, and follows from rule 1:

$$\sum_{i=1}^n (X + c) = \sum_{i=1}^n X + \sum_{i=1}^n c = \sum_{i=1}^n X + nc$$

which provides us with the next rule.

**A.3.4 Rule 4:**  $\sum_{i=1}^n c = nc$

The sum of a constant is simply  $n$  times that constant.

**A.3.5 Rule 5:**  $\sum cX = c \sum X$

What if every number in  $X$  instead of having a constant *added* to it, was *multiplied* by a constant (where the constant has either positive—multiplication— or negative—division—exponents; see section A.2.1)? Again, the original  $\sum X$  could be ignored, each  $X_i$  multiplied by  $c$  to produce a new value,  $Z_i = cX_i$ , and then  $\sum Z$  computed. However, as every  $X_i$  is multiplied by the constant,

$$\sum_{i=1}^n cX_i = cX_1 + cX_2 + \cdots + cX_n$$

can be re-written as:

$$\sum_{i=1}^n cX_i = c(X_1 + X_2 + \cdots + X_n) = c \sum_{i=1}^n X_i$$

which is just the original sum multiplied by  $c$ .

## A.4 Questions

1. The sum of 20 scores is found to be 50. If 2 were to be subtracted from every one of the 20 scores, what would the new sum be?
2.  $X = \{5, 17, 2, 4\}$  and  $\sum_{i=1}^4 Y = 10$ . If each item of  $Y$  were subtracted from its corresponding item in  $X$ , what would the new sum be?
3.  $X = \{2, 11, 5, 17, 2, 8, 9, -4, 2.3\}$ .  $\sum_{i=3}^8 X = ?$ .
4. Each of 5 scores is doubled, and then summed to produce a total of 50. What is the sum of the scores *before* doubling?

## Appendix B

# Keppel's System for Complex ANOVAs

This chapter is loosely based on Appendix C-4 from Keppel (1982), and represents an attempt to render it intelligible. It is included here at the request of my (JRV) senior and former students, who believe, apparently, that only this chapter makes anything other than simple one-way ANOVAs make sense. It is, as one might say, a “good trick”: if you want to wow your colleagues on your ability quickly to determine the appropriate error-terms for the effects of *any* ANOVA design, this appendix provides the algorithm you need.<sup>1</sup> Indeed, I (JRV) have used it as the basis of more than one general ANOVA computer program. Note that where it provides for *no* error-term, a *quasi-F* may be computed, by combining the existing means-squares either strictly by addition ( $F'$ ) or by some combination of addition and subtraction ( $F''$ )—allowing for the unfortunate possibility of negative  $F$ -ratios). The logic of quasi-Fs is, in one sense, unassailable (as long as one is thinking “approximation to a randomization or permutation test”), but is otherwise questionable, at best.

To provide a concrete context, the following research design will serve to exemplify the steps and the logic. Suppose we have 20 nouns and 20 verbs—presumed to be a random sample from the larger population of nouns and verbs. Following the convention of using uppercase letters to denote *fixed* factors and lowercase letters to denote *random* factors, noun vs. verb forms a fixed, word-type factor,  $T$ . Each participant receives either the 20 nouns or the 20 verbs, with 10 participants in each condition. Thus, the random factor, words,  $w$ , is nested within  $T$ , and participants,  $s$ , are also nested within  $T$ , but within each level of  $T$ ,  $s$  crosses  $w$ , as shown in Table B.1.

---

<sup>1</sup>Keppel (1982) is the best book on ANOVA ever written, in my opinion; the next edition (the third) of this book, for reasons I can't fathom, stripped out most of what made Keppel (1982) so good, including Appendix C-4. If you can find a copy of Keppel (1982), hang on to it.

		Word Type ( $T$ )										
		Noun					Verb					
		$w_1$	$w_2$	$w_3$	$\cdots$	$w_{20}$						
							$w_{21}$	$w_{22}$	$w_{23}$	$\cdots$	$w_{40}$	
$s_1$							$s_{11}$					
$s_2$							$s_{12}$					
$s_3$							$s_{13}$					
$\vdots$							$\vdots$					
$s_{10}$							$s_{20}$					

Table B.1: Hypothetical research design.

## B.1 Identifying Sources of Variation

1. List all factors in the study, using a single letter for each, including the experimental unit, as if they all cross. For the example design, these would be:

$$T, w, s$$

2. Construct all possible interactions, again treating the factors as if they all cross:

$$T, w, s, T \cdot w, T \cdot s, w \cdot s, T \cdot w \cdot s$$

3. Denote nesting for each potential source wherever it occurs using the slash notation of Lindman (1974). Where letters repeat to the *right* of the slash, write them only once:

$$T, w/T, s/T, T \cdot w/T, T \cdot s/T, w \cdot s/T, T \cdot w \cdot s/T$$

4. Delete each of the *impossible* interactions—those sources involving nested factors, which will be those in which the same letter occurs on both sides of the slash. For the example, we are left with the following as sources of variance within the design:

$$T, w/T, s/T, w \cdot s/T$$

## B.2 Specifying Degrees of Freedom

Each factor in the design has a specific number of *levels*. For nested main effects, the number of levels is computed *as if* the effect crossed all other effects. Thus, for the example design,  $T$  has 2 levels,  $w$  has 20, and  $s$  has 10.

To compute the degrees of freedom (*df*) for each source, the following two rules are used:

1. If the source contains *no* nested factors, then multiply the *dfs* associated with the different factors listed in the effect.

2. If the source contains a nested factor, then multiply the *product* of the *dfs* of the factors to the left of the slash by the *product* of the *levels* of the factors to right of the slash.

Using lowercase letters of each of the factors to denote the *levels* of each factor, we obtain the following for the sources remaining at the end of step 4 from section B.1:

Source	Degrees of Freedom	Example values
$T$	$(t - 1)$	$(2 - 1) = 1$
$w/T$	$(w - 1)t$	$(20 - 1)2 = 38$
$s/T$	$(s - 1)t$	$(10 - 1)2 = 18$
$w \cdot s/T$	$(w - 1)(s - 1)t$	$(10 - 1)(20 - 1)2 = 342$

Table B.2: Calculation of the degrees of freedom for the research design in Table B.1.

### B.3 Generating Expected Mean Squares

As before, uppercase letters represent fixed factors and lowercase letters represent random factors. Note that a source is considered fixed only if *none* the letters (factors) within it are lowercase or random. Said differently, a source is considered fixed only if *all* the letters (factors) within it are uppercase or fixed. It is otherwise *random*. Each component variance term is multiplied by a set of coefficients corresponding to the number of times the component variance occurs in the design. These coefficients are represented by lowercase letters in parentheses (so that they are not confused with the letters of random factors), and correspond to the levels of the factors *not* in the effect. The product of these levels is the number of observations in the design contributing to the deviations upon which the variance component is based. Thus, for the components in our example, the component variances, complete with coefficients (in parentheses) would be:

$$\begin{aligned}
 T &\Rightarrow (w)(s)T \\
 w/T &\Rightarrow (s)w/T \\
 s/T &\Rightarrow (w)s/T \\
 w \cdot s/T &\Rightarrow w \cdot s/T
 \end{aligned}$$

There are two rules for generating the expected mean squares, in both cases *ignoring the letters corresponding to coefficients expressed in parentheses*:

1. List the variance component term indentifying the source in question. This term is known as the *null hypothesis component* for that source.
2. List additional terms whenever they

- (a) include *all* of the letters defining the source
- (b) such that *all* other letters to the *left* of the slash are lowercase (i.e., random factors). For this purpose, all letters are assumed to be to the left of the slash for non-nested sources.

The expected mean squares for the example design are shown in Table B.3.

Source	Expected Mean Square
$T$	$(w)(s)T + (s)w/T + (w)s/T + w \cdot s/T$
$w/T$	$(s)w/T + w \cdot s/T$
$s/T$	$(w)s/T + w \cdot s/T$
$w \cdot s/T$	$w \cdot s/T$

Table B.3: The expected mean squares for the sources of variance in the design shown in Table B.1.

## B.4 Selecting Error-Terms

Once the expected mean squares have been determined for each source, error-terms for constructing  $F$ -ratios can be selected. The basic notion is simple: form a ratio in which the numerator (the expected mean square of the source to be tested) differs only in the *null hypothesis component* from the denominator. That denominator is the *error-term* for that source. The actual  $F$ -ratio is formed by taking the ratio of the corresponding mean squares. For example, for the effect of “word within Type”,  $w/T$ , the  $F$ -ratio would be:

$$F(df_{w/T}, df_{w \cdot s/T}) = \frac{MS_{w/T}}{MS_{w \cdot s/T}}$$

because the expected mean square for the  $w \cdot s/T$  interaction differs *only* by the null hypothesis component of  $w/T$  from the expected mean square of  $w/T$ , as may be seen in Table B.3. The same is true for the effect of “subjects within Type”,  $s/T$ , and so it takes the same error-term.

But what of the effect of word-Type,  $T$ ? There is no corresponding expected mean-square from among the other effects that matches that of  $T$  in all but the null hypothesis component of  $T$ . With no corresponding error-term, there is no  $F$ -ratio that may be formed to test the effect of word-Type,  $T$ .

## Appendix C

# Answers to Selected Questions

Yes, as students, we always hated the “selected” part, too. Until, that is, we wrote our own book to accompany the course. Why not answer *every* question, so the reader can see how the “expert” (or, at least, the textbook authors) would answer each and every question? The concern is just that, because the answers would thereby be *so* readily available, the reader would not take the extra minutes to solve the problem on her or his own. And it is those extra few minutes of deliberation and problem solving that we believe are essential to gaining a good grasp and feel for the material. “Fine”, you might say. “But is not watching the ‘expert’ at work valuable, and, indeed, *instructive*?” We can’t but agree. View it as a compromise. Frustrating? Yes. Necessary? Also yes.

Rather than simply provide each numeric answer itself, we have worked out each of the answers here in some detail so it should be apparent *how* we arrived at each of them. We urge the reader to do the same for the remaining unanswered questions.

### C.1 Chapter 1: Page 5

1. The arithmetic average is computed as the sum of the scores *divided* by the number of scores, whereas we compute the sum of a (constant) series by *multiplying* one-half of the sum of the highest score and lowest score by the number of scores,  $n$ . So, as dividing by  $n$  to compute the mean simply cancels the multiplication by  $n$ , the mean of a series must be one-half of the sum of the highest and lowest scores, or, in this case,  $(11 + 75)/2 = 43$ .

## C.2 Chapter 2: Page 21

2. By definition the mean is that value such that the sum of deviations from it is equal to 0. Hence,  $q$  must equal the mean, or 11.
5. If Allen Scott had the lowest grade of the 5 males in the class who otherwise would have had the same mean of 90 as the 10 females, then the remaining 4 males must have had a mean 90, for a *sum* of  $4 * 90 = 360$ . The sum for the 10 females, then, was  $10 * 90 = 900$ , and for the class as a whole,  $15 * 85 = 1275$ . Thus, Allen Scott's grade must have been the difference between the sum for the class as a whole minus the sums for the females and the 4 males or  $1275 - 900 - 360 = 15$ .

## C.3 Chapter 3: Page 33

1. As shown in equation 3.6 on page 30, the variance can be calculated as the difference between the mean-square and the squared mean. In the current case, that equals  $S^2 = 2500/20 - 10^2 = 125 - 100 = 25$ . The standard deviation,  $S$ , is therefore  $S = \sqrt{25} = 5$ .

## C.4 Chapter 4: Page 40

1. As detailed on page 39 of section 4.2.1, the sum of squared  $Z$ -scores is equal to  $n$ . Hence, as the sum is 10,  $n = 10$ .
3. A score of 22 from a distribution with a mean of 28 and a standard deviation of 3 is  $(22 - 28)/3 = -2$  standard deviations from the mean. We want it to be the same -2 standard deviations from the mean in the new distribution. As this distribution has a mean of -10 and a standard deviation of .25, then a score of  $-10 + (-2)(.25) = -10.5$  would be 2 standard deviations below the mean.
4. If a score of 50 (Bryan Ian) corresponds to a standard score of -3 and a score of 95 (Margaret Jean) corresponds to a standard score of 2, then the range of  $95 - 50 = 45$  corresponds to a standard score range of  $2 - (-3) = 5$  standard deviations. Thus, if a range of 45 corresponds to 5 standard deviations, then 1 standard deviation corresponds to  $45/5 = 9$  units in the original distribution. Hence, the mean is  $50 + 3(9) = 95 - 2(9) = 77$ .
6. Your dog's score of 74 on the meat lover's subscale corresponds to a WAIS score of 70, which is 2 standard deviations ( $2 * 15 = 30$ ) below the WAIS mean of 100. Therefore, 74 must be 2 standard deviations below its mean. As the standard deviation of the sub-scale is 7, then for 74 to be 2 standard deviations below the mean, the mean must be  $74 + 2(7) = 88$ .

## C.5 Chapter 5: Page 45

2. It would be more peaked (i.e., relatively *light* tails), with more scores in the *right* tail than in the left tail.

## C.6 Chapter 6: Page 51

1. Obviously, with a mean of -4, the distribution does *not* meet the requirements (i.e., no negative scores) of Markov's inequality theorem; thus, based on the Tchebycheff inequality, scores of 6 are  $(6 - (-4))/5 = 2$  standard deviations from the mean, and, hence, at most  $1/(2^2) = .25$  or  $.25(50) = 12.5$  of the scores could be beyond plus or minus 2 standard deviations from the mean.

## C.7 Chapter 7: Page 63

3. 80% (16/20) of the students were psychology majors, and 20% were business majors. From section 7.3.2, page 61 of Chapter 7, we compute the point-biserial correlation coefficient:

$$r_{pb} = \frac{80 - 76}{2} \sqrt{(.8)(.2)}$$

$$r_{pb} = 0.8$$

Thus, a substantial correlation of major with sadistics (sic) grade, with psychology majors exceeding business majors.

## C.8 Chapter 8: Page 74

5. The standard deviation of the *predicted* scores is given by  $S_{Y'} = S_Y(r) = 10(.8) = 8$ .

## C.9 Chapter 9: Page 87

3. Because we want to take sex into account for both variables, a *partial* correlation is called for. Thus,

$$r_{(gb).s} = \frac{-.83 - (-.8325)(.665)}{\sqrt{1 - (-.8325)^2} \sqrt{1 - .665^2}}$$

$$r_{(gb).s} = \frac{-.2763}{.41} = -.67$$

**C.10 Chapter 10: Page 98**

1. The product can be positive only if (a) all 4 numbers are positive, (b) all 4 numbers are negative, or (c) 2 numbers are positive and 2 are negative. There are  $\binom{8}{4}\binom{6}{0}$  ways the numbers can all be positive,  $\binom{8}{0}\binom{6}{4}$  ways the numbers can all be negative, and  $\binom{8}{2}\binom{6}{2}$  ways 2 of the numbers can be positive and 2 negative. There are  $\binom{14}{4}$  total ways of selecting 4 numbers from the 14, so the answer is:

$$\frac{\binom{8}{4} + \binom{6}{4} + \binom{8}{2}\binom{6}{2}}{\binom{14}{4}}$$

Some people are not happy leaving the answer in this form, feeling, apparently, that if the answer is not reduced to a *number* the question has not been fully answered. To that end, we'll get you started by noting that the term  $\binom{6}{4}$  is  $\frac{6!}{4!2!}$ , which is  $\frac{(6)(5)(4!)}{2!4!}$ . Cancelling the 4!, yields  $\frac{(6)(5)}{2} = (3)(5) = 15$ .

**C.11 Chapter 11: Page 107**

2. The probability of getting 10 or more heads in 12 tosses of a fair (i.e.,  $p = .5$ ) coin is:

$$\binom{12}{10}.5^{10}.5^2 + \binom{12}{11}.5^{11}.5^1 + \binom{12}{12}.5^{12}.5^0 = .0193$$

As this is less than .05, it is not likely to be a fair coin.

- 6a. The answer is one minus the probability that fewer than three have taken statistics:

$$1 - \left( \binom{10}{0}.2^0.8^{10} + \binom{10}{1}.2^1.8^9 + \binom{10}{2}.2^2.8^8 \right)$$

**C.12 Chapter 12: Page 125**

1. Because there are more than 30 scores, we can assume that the sampling distribution of *means* will be roughly normally distributed. The standard error of the mean is  $18/\sqrt{(36)} = 3$ . So, a  $\mu$  of 7 is 1 standard deviation *below* the sample mean of 10. Thus, a  $\mu$  of less than 7 should give rise to a sample mean of 10 at most 15.87% of the time.
11. Again, because there are more than 30 scores, we can assume that the sampling distribution of *means* will be roughly normally distributed. The standard deviation of a uniform distribution is given by (see section 12.3.3):

$$\sqrt{\frac{(b-a)^2}{12}}$$

So, for the current case,  $\sqrt{(\sqrt{12}(8))^2/12} = 8$ . The standard error of the mean, then, is  $8/\sqrt{64} = 8/8 = 1$ . The mean of a uniform distribution is simply  $(b - a)/2 = \sqrt{12}(8)/2 = 13.86$ . Thus, as 12 is  $12 - 13.86 = -1.86$  or 1.86 standard deviations below the mean, the probability that the mean of the sample is greater than 12 is  $.5 + .4686 = .9686$

### C.13 Chapter 13: Page 135

2. Based on at most 1/3 liking sadistics, the expected value in a sample of 36 ( $\geq 30$ , so the sampling distribution may be assumed to be roughly normal) former students liking sadistics is  $36/3 = 12$ , with the remaining 24 disliking sadistics. However, 16 of the sample actually liked sadistics, so  $z^2 = (16 - 12)^2/12 + (20 - 24)^2/24 = 2$ . Comparing that value to the  $\chi^2$  table with 1 degree of freedom, we find that the probability of observing such a result given the truth of the null hypothesis is between .15 and .20—much greater than the .05 usually required for significance. Hence, we conclude that there is not sufficient evidence to reject the hypothesis that the sample is a random sample of students from the general population of students, at least with respect to liking or disliking sadistics.
- 6b. In the table, 35 ordered beer and 25 didn't. Of the 60, 2/3 or 40 would be expected to order beer. So,  $z^2 = (35 - 40)^2/40 + (25 - 20)^2/20 = 1.875$ . Comparing that value to the  $\chi^2$  table with 1 degree of freedom, we find that the probability of observing such a result given the truth of the null hypothesis is between .15 and .20—much greater than the .05 usually required for significance. Hence, we conclude that there is not sufficient evidence to reject the hypothesis that the sample is a random sample of students from the general population of students, at least with respect to ordering beer.

### C.14 Chapter 14: Page 148

6. The mean of the 5 distances is 258 km, and the estimated standard error of the mean equals 1. Computing  $t_4 = \frac{258 - 260}{1} = 2$ . Looking that up in the  $t$ -table with 4 degrees of freedom, we find that a  $t_4 = 2$  would occur with a probability of between .1 and .2 (two-tailed test), according to the null hypothesis. Hence, there is *not* sufficient evidence to reject the belief that the distance between your home and your friend's over this route is 260 km.

### C.15 Chapter 15: Page 156

- 1b. As the  $F$ -ratio has 3 and 12 degrees of freedom, and the groups were all the same size, there must have been  $3 + 1 = 4$  groups with  $12/4 = 3$  degrees of

freedom per group, meaning there were  $3 + 1 = 4$  observations per group, for a total of  $4(4) = 16$  observations. Alternatively, as  $n = df_{total} + 1$ , then  $n = 3 + 12 + 1 = 16$ .

## C.16 Appendix A: Page 180

4. If 50 equals the sum of the scores *after* doubling, then it equals twice the original sum (because every score was doubled). Hence, the original sum would be  $50/2 = 25$ .

# References

- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, D.C.: Author.
- Begg, I. (1987). Some. *Canadian Journal of Psychology*, *41*, 62–73.
- Bell, E. T. (1937). *Men of mathematics*. New York: Simon and Schuster.
- Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Marcel Decker.
- Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, New Jersey: Prentice-Hall, Inc.
- Lindman, H. R. (1974). *Analysis of variance in complex experimental designs*. San Francisco: W. H. Freeman.
- Rouanet, H., Bernard, J., & Lecoutre, B. (1986). Nonprobabilistic statistical inference: A set-theoretic approach. *The American Statistician*, *40*, 60–65.
- Salsburg, D. (2001). *The lady tasting tea*. New York, New York: W. H. Freeman and Company.
- Shah, A. K. (1985). A simpler approximation for areas under the standard normal curve. *The American Statistician*, *39*, 80.
- Spicer, C. C. (1972). Algorithm AS 52. Calculation of power sums of deviations about the mean. *Applied Statistics*, *21*, 226–227.
- Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, Massachusetts: Harvard University Press.
- Vokey, J. R. (1990). Multiple linear regression and the recursive-SSCP and sweep-regression algorithms. *MicroPsych Network*, *4*, 73–84.
- Vokey, J. R. (1998). Statistics without probability: Significance testing as typicality and exchangeability in data analysis. *Behavioral and Brain Sciences*, *21*, 225–226.

# Index

- Absolute, 35
- absolute mean, 16
- absolute value, 16
- affine transform, 35
- $\alpha$ , 104
- alternative hypothesis, 105
- American Psychological Association, APA, 165
- American Psychological Society, APS, 165
- analysis of variance, 151
- ANOVA, 151
- APA format, 165
- arithmetic average, 9
- ASD, 57
- atypicality, 93
  
- balancing-point, 12, 47
- Ballantine's, 83
- Bartels, Johann Martin, 2
- bell-curve, 44
- Bernoulli process, 102
- Bessel, F. W., 124
- $\beta$ , 105
- biased estimates, 142
- binary, 101
- binomial, 101
- binomial distribution, 101
- binomial standard deviation, 119
- binomial variance, 119
- bivariate normal, 54
- British Psychological Association, BPA, 165
  
- Calgary, 27
- Cambridge, England, 91
  
- Canadian Psychological Association, CPA, 165
- Canadian Society for Brain, Behaviour, and Cognitive Science, CS-BBCS, 165
- Catbert, 76
- categorical association, 62
- Celsius, 35
- central limit theorem, 111
- centred distribution, 12
- centring the distribution, 12
- chi square distribution, 127
- $\chi^2$  distributions, 128
- $\chi^2$  tests of independence, 132
- chi-square, 62
- coefficient of alienation, 73
- coefficient of determination, 73
- coefficient of non-determination, 73
- Coke, 107, 132
- combinations, 95
- complementary event, 49
- completely-randomised designs, 151
- computational formulae, 30
- confidence interval, 122, 123
- constant series with  $c > 1$ , 3
- contingency tables, 2 x 2, 132
- contradictories, 101
- contraries, 101
- controversy, 92
- conventional wisdom, 97
- correlation, 53
- correlation coefficient, 53
- correlation matrix, 82
- Covariance, 58
- $\phi_C$  (Cramér's  $\phi$ ), 135

- $D^2$ -statistic, 26
- $D$ -statistic, 26
- De Moivre, Abraham, 111
- Definition of a constant series, 3
- degrees of freedom, 128, 133, 142, 153
- deviation scores, 12
- Edgeworth, F. Y., 53
- error of prediction, 67
- estimated population variance, 143
- estimated standard error of the difference, 146
- estimated standard error of the mean, 143
- Euler diagrams, 83
- Euler, Leonhard, 83
- exchangeability, 97
- exhaustive, 101
- expected frequencies, 128
- experimental design, 91
- Experimental Psychology Society, EPS, 165
- F-ratio, 155
- factor analysis, 82
- Fahrenheit, 35
- first moment, 43
- Fisher, Sir Ronald A., 91
- floating-point, 30
- fourth moment, 44
- fundamental counting rule, 94
- Galbraith, John Kenneth, 97
- Galton, 45
- Galton, Sir Francis, 53, 65
- Gauss, Carl Friedrich, 111
- Gauss, Sir Carl Friedrich, 1
- Gaussian distribution, 44, 111
- geometric mean, 17
- gila monsters, 63
- goodness of fit, 127
- Gosset, W. S., 139
- grading on a curve, 35
- GRE scores, 35
- GRE—Graduate Record Exam, 39
- Guinness breweries, 139
- hand calculation, 30
- harmonic mean, 18
- homogeneity of variance, 152
- horse races, 64
- independence, 93
- independent, 101
- IQ, 44
- IQ scores, 35
- IQ tests, 40
- Kamloops, 27
- kurtosis, 44
- Lamborghini Countach, 149
- Laplace, 111, 124
- least-squares criterion, 67
- Legendre, 111
- leptokurtic, 44
- linear intensity, 65
- linear regression, 65
- linear transform, 35
- linear transformation, 65
- lung cancer and smoking, 81
- Markov inequality, 47
- Markov, Andrei Andreyevich, 47
- matrix algebra, 86
- mean, 9
- mean square, 29
- mean squares, 153
- median, 10, 19
- Medicine Hat, 27
- mesokurtic, 44
- method of provisional means, 15
- moments of distributions, 43
- multiple correlation, 62, 82
- mutually exclusive, 101
- Neyman, Jerzy, 122
- normal distribution, 44, 111
- normalised sum, 9
- null hypothesis, 105
- numeric precision, 30

- observed frequencies, 128
- one-tailed test, 104
- parabola, 12
- partial correlation, 79
- partitioning variance, 73
- Pearson product-moment correlation coefficient, 134
- Pearson, Egon, 122
- Pearson, Karl, 53, 65, 127
- Pepsi, 107, 132
- Pepsi Challenge, 107
- permutation and exchangeability tests, 151
- permutations, 95
- phi coefficient, 62
- $\phi$  (phi coefficient), 134
- $\pi$ , 93
- PIP, 73
- platykurtic, 45
- platypus, 45
- Point-Biserial Correlation Coefficient, 61
- polytomous variables, 62
- population distribution, 102
- population parameters, 92, 105
- power, 105
- principal component, 68
- principal components analysis, 82
- probability, 92
- probability density function (pdf), 111
- pseudo-random, 93
- psychology, 93
- $r$  (correlation coefficient), 53
- $R$  (multiple correlation), 86
- $r \times c$  contingency tables, 134
- random assignment, 91, 107
- random sampling, 92, 93, 107
- random sequence, 93
- randomisation, 93
- randomisation testing, 151
- range, 25
- recursive computation, 16
- reductio ad absurdum*, 12
- regression, 53
- regression coefficient, 66
- regression equation, 67
- regression line, 66
- rescaling, 35
- residual sum-of-squares, 69
- reversion, 53
- root-mean-square, 17
- running mean algorithm, 15
- $S$  (standard deviation), 26
- $S^2$  (variance), 26
- sadistics, 63
- sampling distribution, 102
- sampling with replacement, 101
- SAT scores, 35
- scientific psychology, 165
- second moment, 43
- semi-partial correlation, 80
- sign test, 107
- significance, 91
- skewness, 43
- slope, 65
- small-sample approach, 139
- Smith, H. Fairfield, 91
- Spearman Rank-Order Correlation Coefficient, 60
- squared deviations, 12
- standard deviation, 26
- standard deviation (definition), 29
- standard error of estimate, 69
- standard error of the mean, 121
- standard score transform, 38
- standard-score, 38
- statistic, 10
- statistical significance, 91
- structure, 53, 91
- Student, 142
- sum of squares, 12
- sum of squares between groups, 153
- sum of squares within groups, 153
- Summing any Constant Series, 3
- Sums of Cross-Products, 54
- Sums of Differences, 56
- Swift Current, 27

- t-distributions, 142
- t-test, 143
- t-test for dependent samples, 145
- t-test for independent samples, 146
- tautology, 152
- Tchebycheff inequality, 50
- Tchebycheff, Pafnuti L., 50
- Tied Ranks, 61
- Trans-Canada highway, 27
- transformed scores, 35
- two-tailed test, 103
- Type I error, 103
- Type II error, 103
  
- unbiased estimates, 142
- uniform distribution, 45, 121
- uniform distribution, mean, 121
- uniform distribution, standard deviation, 121
  
- Vancouver, 27
- variability, 25
- variance, 26
- variance (definition), 29
  
- WAIS, 40
- Wechsler Adult Intelligence Scale, 40
- Weldon, Walter, 53
- Wesayso corporation, 76
- which hunt, 170
  
- Z-score, 38