Chapter 1

INTRODUCTION

The correlation matrix is a row-by-column arrangement of a set of correlation coefficients. The rows and columns refer to specific variables, which are measured features of the people, animals, or entities that behavioral science researchers study. For example, four variables assessed on people may be height, intelligence, birthweight, and shyness; three variables assessed on animals may be cortisol levels, reaction time, and counts of observed behaviors; and three variables assessed on an entity (e.g., a school) may be the percent low income, teacher turnover rate, and average student performance. A correlation matrix indicates the linear association between each pair of variables, such that the same variables in the same order label both the columns and the rows of the correlation matrix.

However, a correlation matrix is much more than an arrangement of individual correlation coefficients. Dozens of careful treatments of the correlation coefficient itself—the elements of a correlation matrix—exist in the statistical literature (e.g., Chen & Popovich, 2002; Rodgers & Nicewander, 1988) and in both sophisticated and introductory statistics textbooks. But understanding and appreciating the correlation *matrix* requires rather more careful study and mathematical sophistication than is required to understand the correlation *coefficient*. Few treatments—at either the introductory or more advanced level—extend the pedagogy of correlations from the separate correlation coefficients to the overall integrated correlation matrix. The current book, directed toward students, researchers, and methodologists who need to understand and/or teach correlation matrices, aims to provide this treatment.

We begin with a brief review of the correlation coefficient and of the related measure, the covariance. Correlation and covariance provide the foundation for many statistical techniques used across social, behavioral, and biological science disciplines. They also appear often in engineering, medical research, operations research, the physical sciences such as physics and chemistry, and other disciplines. Because correlations and covariances are the starting points for many statistical procedures, any discipline that defines its methods through statistical analysis is likely to rely extensively on these two measures of relationship. We treat both correlations and covariances throughout, though we will emphasize the correlation and, thus, will typically refer only to the correlation in general treatment. We will make clear when we are treating one or the other specifically. We distinguish the correlation and the covariance later in this chapter.

The Correlation Coefficient: A Conceptual Introduction

The correlation coefficient describes the linear association between two variables. It answers the question, "When one variable decreases or increases, how does the other variable tend to decrease or increase?" Correlation coefficients range from −1 to +1; magnitudes greater in absolute value (closer to +1 or −1) indicate a stronger association. Positive values indicate that as one variable increases (decreases), the other variable tends to increase (decrease)—that is, a positive or direct relationship. Negative values indicate that as one variable increases, the other variable tends to decrease (and vice versa)—that is, a negative or inverse relationship.

There are a number of different types of correlation coefficients, each with the purpose of quantifying the linear relationship between two variables. One way to categorize correlation coefficients into a taxonomy is defined based on the measurement level (Stevens, 1946) of the variables involved. That taxonomy assesses whether one or both variables are categorical (nominal), ordinal, or quantitative (interval or ratio). Often, "correlation" is shorthand for the Pearson product–moment correlation coefficient, the most common correlation measure that is used when both variables are quantitative, measured at interval or ratio levels within the Stevens measurement level system. Other correlation coefficients have been defined as well. For two ordinal variables, Spearman's rho, Kendall's tau, and polychoric correlation could be used to measure the relationship. For two binary (dichotomous) outcome variables, the phi coefficient and the tetrachoric correlation are the appropriate measures of association. For one binary and one quantitative variable, the point-biserial and biserial correlation coefficients are appropriate correlation measures.

Another classification system is the one used by Chen and Popovich (2002), which distinguishes between parametric and nonparametric measures. This distinction typically involves the question of whether a normal distribution is assumed to underlie one or both variables. For example, the formulas for polychoric and tetrachoric correlations assume that a normal distribution underlies the nonquantitative variables of interest. On the other hand, a nonparametric correlation "requires fewer assumptions and does not attempt to estimate population parameters" (Chen & Popovich, 2002, p. 79). Many nonparametric correlations are computed from variables that are naturally categorical (e.g., bright vs. dark colors, urban vs. rural residence) and do not have any underlying quantitative distribution, normal or otherwise. The phi coefficient and Spearman's rho are examples of nonparametric correlations.

A third way to classify correlations is in relation to the original correlation measure, the Pearson correlation. Several of the correlations defined above are actually special cases of the Pearson correlation applied to nonquantitative variables. For example, given two ordinal variables, we can rank order their values (within each variable); if we apply the Pearson correlation formula to those rank orders, we compute a Spearman rho correlation coefficient. Similarly, suppose we have two variables, one a typical quantitative variable and the other a binary variable coded with 0 indicating one category and 1 the other. If we use the Pearson correlation formula on that coding scheme, we are computing a point-biserial correlation coefficient. The phi coefficient is also a special case of the Pearson correlation, defined using the Pearson formula for two variables, each measured as binary variables. On the other hand, the Kendall tau ordinal correlation, the biserial correlation, and the tetrachoric/polychoric correlations are defined using different formulas that are not special cases of the Pearson correlation.

There exist many different ways to interpret a correlation coefficient. Rodgers and Nicewander (1988) showed that the correlation coefficient can be interpreted as one of several special kinds of means (e.g., the mean of the standardized cross products, or as a geometric mean), a special case of covariance, a special kind of variance, the slope of the standardized regression lines, a cosine, a function of the angle between two regression lines, and through several additional trigonometric interpretations. Standard introductory statistical textbooks show how to do null hypothesis significance testing (NHST) using the correlation coefficient, and the correlation measure can also be interpreted as an effect size. To conclude our review of the correlation coefficient—the building block for the correlation matrix—we present one of the standard formulas for the correlation coefficient. There exist many formulas that are algebraically equivalent, but conceptually distinct. We use one that allows us easily to demonstrate its relationship to a measure of covariance. If we define two quantitative variables, X and Y, with means $\bar{X}$ and $\bar{Y}$, respectively, for N observations (i.e., we define N pairs of scores), then the Pearson correlation coefficient can be computed using the following formula:

$$r_{XY} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\,\Sigma(Y-\bar{Y})^2}} \tag{1.1}$$

The Covariance

Earlier, we referred to the covariance as a measure similar to the correlation. The cleanest way to conceptualize the relationship between the correlation and the covariance is to consider the correlation as a standardized version of the covariance. In other words, the correlation can be viewed as a measure of relationship between standardized variables, whereas the covariance is the measure of relationship between the equivalent unstandardized (or raw) variables. To appreciate this distinction requires understanding unstandardized and standardized variables.

When scores on a variable are collected using a particular scale of measurement (e.g., intelligence quotient [IQ], with a mean of 100 and standard deviation of 16; a shyness scale, with a mean of 50 and a standard deviation of 5; or adult female height, with a mean of 65 inches and a standard deviation of 3 inches), we typically refer to those measures as raw scores, measured on an unstandardized variable. If a respondent has a score of 68 inches on the height scale, and a score of 92 on the IQ scale, it is meaningless to compare those two raw scores; in no sense does the respondent have 24 more units of IQ than of height, because of the different scales of measurement.

This incompatibility is easily adjusted using standardization. Standardized scores (also called *z* scores) are defined for a given variable by subtracting the variable's mean and dividing by the variable's standard deviation; standardized scores indicate how far above or below the variable's mean that score is in terms of standard deviation (*SD*) units. Thus, the standardized score associated with a height of 68 inches is $z_{\text{height}} = (68 - 65)/3 = 1$; this computation tells us that a height of 68 is 1 *SD* unit above the mean. The standardized score associated with an IQ of 92 is $z_{\text{IQ}} = (92 - 100)/16 = -0.5$; this computation tells us that an IQ of 92 is 0.5 *SD* units below the mean. At this point, these two measurement scales have been standardized and are now at least loosely comparable.

This development allows us to distinguish the correlation from the covariance. The correlation can be defined in relation to *z* scores. A mathematically equivalent form of the Pearson correlation formula in Equation (1.1) is the following formula:

$$r_{XY} = \frac{\sum z_X z_Y}{N-1} \tag{1.2}$$

Furthermore, the correlation has defined bounds of +1.0 and −1.0. The covariance, which has no defined bounds in general, depends on the scales of measurement of the two variables. The formula for the covariance between two variables, *X* and *Y*, is the following. Note the similarity between Equations (1.2) and (1.3).

$$\text{Cov}(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N-1} \tag{1.3}$$

In some research settings, it is important to define statistical procedures that account for the different scales of measurement of the variables. We will briefly touch on such issues when we discuss factor analysis and structural equation modeling (SEM) in Chapter 4. In such settings, covariances—and covariance matrices—are the preferred measures of association. In other settings, the researcher would prefer to equate the scales of measurement—using standardization—so that differences in the scales' means and standard deviations can be ignored. In those settings, correlations—and correlation matrices—are the preferred measures of association. There is no correct answer to the question, "Which should be used, a correlation (matrix) or a covariance (matrix)?" The answer depends on the researchers' goals and how the variables were measured. This book—ostensibly about correlation matrices—is also about covariance matrices as well. By the end of the book, the reader will have some insight into when correlation matrices are preferable to covariance matrices, and vice versa. We emphasize, however, that our typical (and default) treatment in this book is of the correlation matrix.

The Correlation Coefficient and Linear Algebra: Brief Histories

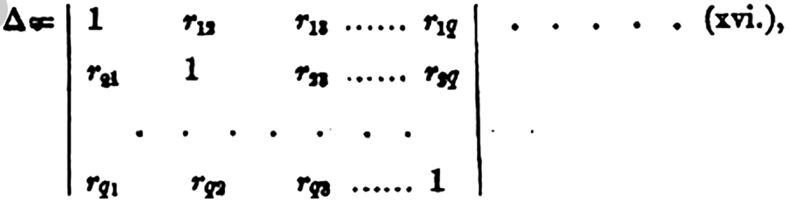
It is not coincidence that the two developers of the correlation coefficient—Francis Galton and Karl Pearson, in the late 1800s—were collectively interested in a wide range of scientific disciplines, including psychology, genetics, geography, astronomy, sociology, and biometrics (Stanton, 2001). These fields required a measure that would appropriately capture the association between two quantitative variables. Galton first proposed the idea of the correlation coefficient, stemming from his earlier work on regression (Galton, 1885), while conducting research on the correspondence between parents' and their offspring's physical traits. Through this work, he realized there existed an "index of correlation" that captured the linear association between heights in kinship pairs. By 1890, he understood that the idea of correlation extended beyond questions of heredity and could be applied broadly to any two quantitative variables—and not simply to measures of the same construct, as he had originally thought (Stigler, 1989). However, it was his student, Pearson, who developed the mathematical formula and theory of the product–moment correlation that is still used most commonly today (Pearson, 1896).

Preceding the development of correlation by only a few decades was the development of linear, or matrix, algebra. Linear algebra grew out of the study of determinants for systems of linear equations in the early 1800s. Determinants are measures obtained from a matrix that reflect the linear relationships inside the matrix and, thus, are mathematically related to correlations. Interestingly, the mathematical concept of determinants (now wedded to the mathematics of matrices) developed well before matrix algebra; determinants were referenced at least as early as the 17th century by Leibnitz. In 1848, J. J. Sylvester first used the term *matrix* in a mathematical setting, the word *matrix* deriving from Latin for "womb," "mother," or "place where something develops." In 1855, Arthur Cayley first referred to a matrix with a single, uppercase letter, thereby cementing matrices as entities more complete than and distinguishable from their separate elements. The first linear algebra textbook, appropriately titled *Linear Algebra* by Hüseyin Tevfik Pasha, was written by happenstance almost contemporaneously with the development of the correlation coefficient (though linear algebra developed in what is today Bulgaria, whereas correlation and regression developed largely in England).

The development of both matrix algebra and the correlation coefficient set the stage for the rapid development of the correlation matrix and statistical methods applied to the correlation matrix. Unsurprisingly, Pearson was one of the first psychometricians to incorporate the newly developed and quickly expanding field of matrix algebra into his conceptualization of the correlation coefficient. In his groundbreaking 1901 article, in which he proposed what would later become principal components analysis (PCA), Pearson demonstrated both the computation of a determinant and what a correlation matrix between *q* variables would look like (see Figure 1.1).

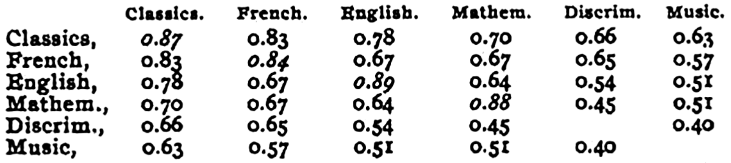
However, he did not refer to the mathematical entity he created as a correlation matrix (the term *matrix* does not appear in the article), nor did he consider the correlation matrix beyond its convenient notation for producing the determinant. Three years later, Spearman (1904), a psychologist who made extensive contributions to statistics (including early work on factor analysis and adapting Pearson's correlation formula for ordinal variables), published what may be the first empirical correlation matrix (except that the diagonal had been modified; Figure 1.2); these correlations relate measures among British schoolchildren of "talent" within these different "branches."

Figure 1.1 The Determinant of a Generic Correlation Matrix, Appearing in Pearson (1901)



$$\Delta = \begin{vmatrix} 1 & r_{12} & r_{13} \ldots\ldots & r_{1q} \\ r_{21} & 1 & r_{23} \ldots\ldots & r_{2q} \\ . & . & . \quad . \quad . & . \\ r_{q1} & r_{q2} & r_{q3} \ldots\ldots & 1 \end{vmatrix} \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(xvi.)},$$

Figure 1.2 A Modified Correlation Matrix, Appearing in Spearman (1904, p. 275)



	Classics.	French.	English.	Mathem.	Discrim.	Music.
Classics,	*0.87*	0.83	0.78	0.70	0.66	0.63
French,	0.83	*0.84*	0.67	0.67	0.65	0.57
English,	0.78	0.67	*0.89*	0.64	0.54	0.51
Mathem.,	0.70	0.67	0.64	*0.88*	0.45	0.51
Discrim.,	0.66	0.65	0.54	0.45		0.40
Music,	0.63	0.57	0.51	0.51	0.40	

Similar to Pearson, Spearman (1904) did not refer to his table as a "correlation matrix" but rather a "table of correlation," with instructions to the reader for how to read the table: "Each number shows the correlation between the faculty vertically above and that horizontally to the left" (p. 274).

Correlation matrices in scholarly literature had a small but consistently increasing number of mentions in the decades following the contributions of Spearman and Pearson. However, in recent decades, the explicit use of correlation matrices has increased exponentially. According to an informal search of the ProQuest online scholarly text database, there were 32 peer-reviewed records in the 1930s that mentioned the term *correlation matrix*. By the 1980s, the number of records grew to 1,827. By the most recent count for the 2010s that number is 31,505, with references in mathematics, neuroimaging, environmental science, applied psychology, and business journals. The wealth of attention to correlation matrices in applied research is likely due to a conflux of multiple factors, including modern computation, advances in data collection techniques, and advances in methods to analyze correlation matrices.

We provide here a brief summary of the history of correlation matrices to support what is well-known among statisticians, but which is less obvious for novice statistical students: Correlation matrices are more than just the convenient square arrangement of correlation coefficients. Inspection of correlation matrices facilitates a deeper understanding of the *multivariate* relationships among variables and allows for more complex theory development and testing than can possibly emerge from inspection of the separate disjoint correlations. One way to begin to appreciate the nuance of a correlation matrix is to recognize that not only does a correlation matrix include information about pairs of variables, but it also implicitly contains mathematical constraints that apply to relationships among triples of variables, to quadruples of variables, and so on. This book is written to develop intuition and understanding for correlation matrices, the tests that may be conducted on them, the modeling that can be applied to them, and the graphical methods that may be used to display them.

Examples of Correlation Matrices

In the following paragraphs, we develop a number of examples of correlation matrices (and, in several cases, the equivalent covariance matrix as well). These examples are based on real data collected in real research settings. They are chosen to be disciplinarily broad, including variables that would be used in education, psychology, sociology, political science, economics, communications, health care research, and other social/behavioral sciences settings. Once defined, we use these specific correlation matrices throughout the chapters of this book to illustrate principles and application of statistical methods relevant to correlation matrices.

As an example of how correlation matrices can motivate hypotheses or empirical analyses that are difficult to interpret using only bivariate correlations, consider Tables 1.1 and 1.2 adapted from Humphreys et al. (1985). Each table represents a correlation matrix capturing how a construct (intelligence of boys and girls, respectively) correlates within person over development. Although inspection of any given element of the correlation matrix would indicate that boys' (or girls') intelligence at one time is positively associated with intelligences at another time, the correlation matrix structure makes salient that intelligence measurements closer together are more strongly correlated than those measured further apart. Furthermore, the patterns of correlations are similar for boys and girls, and using methods presented in this book, we can formally test if the correlations are equivalent across gender in the population. In addition, at younger ages intelligence does not correlate as strongly with adjacent time points as it does at later time points. These kinds of observations would not be either obvious or easy to discuss if we relied on inspection of separate correlation coefficients. Within the context of a correlation matrix, their description and study are straightforward.

Although the correlations in Tables 1.1 and 1.2 were calculated on individuals (i.e., children across years of development), correlation matrices are also frequently used to show relationships for which the unit of analysis is a group. As examples, we have included two correlation matrices based on groups. Table 1.3 demonstrates a correlation matrix, adapted from Elgar (2010), calculated from 33 countries for which the variables of interest are country-level indicators of income inequality, average tendency to trust others, public health expenditures, life expectancy, and adult mortality.

Table 1.1 Correlations Between Boys' Intelligence Measured at Different Ages

Age (Years)	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
8	1.00	.60	.63	.67	.64	.59	.60	.62	.62	.53
9	.60	1.00	.74	.70	.68	.68	.68	.59	.60	.57
10	.63	.74	1.00	.79	.77	.70	.75	.71	.71	.60
11	.67	.70	.79	1.00	.87	.78	.75	.79	.81	.75
12	.64	.68	.77	.87	1.00	.84	.79	.77	.80	.76
13	.59	.68	.70	.78	.84	1.00	.85	.77	.77	.77
14	.60	.68	.75	.75	.79	.85	1.00	.84	.80	.75
15	.62	.59	.71	.79	.77	.77	.84	1.00	.88	.78
16	.62	.60	.71	.81	.80	.77	.80	.88	1.00	.85
17	.53	.57	.60	.75	.76	.77	.75	.78	.85	1.00

Note: The longitudinal sample of boys from the Boston area used a variety of measures of intelligence across the 10 yearly time points. Correlations in the original article were calculated pairwise, with sample sizes differing between 391 and 511; for examples throughout the book, we use a conservative *N* of 391. Adapted from Humphreys et al. (1985).

Table 1.2 Correlations Between Girls' Intelligence Measured at Different Ages

Age (Years)	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
8	1.00	.67	.64	.70	.69	.64	.64	.64	.63	.54
9	.67	1.00	.65	.68	.73	.73	.69	.61	.61	.61
10	.64	.65	1.00	.78	.78	.73	.73	.73	.69	.59
11	.70	.68	.78	1.00	.88	.80	.79	.80	.80	.75
12	.69	.73	.78	.88	1.00	.85	.84	.79	.79	.77
13	.64	.73	.73	.80	.85	1.00	.85	.75	.77	.79
14	.64	.69	.73	.79	.84	.85	1.00	.81	.77	.75
15	.64	.61	.73	.80	.79	.75	.81	1.00	.90	.79
16	.63	.61	.69	.80	.79	.77	.77	.90	1.00	.87
17	.54	.61	.59	.75	.77	.79	.75	.79	.87	1.00

Note: The longitudinal sample of girls from the Boston area used a variety of measures of intelligence across the 10 yearly time points. Correlations in the original article were calculated pairwise, with sample sizes differing between 495 and 693; for examples throughout the book, we use a conservative *N* of 495. Adapted from Humphreys et al. (1985).

Table 1.4 presents correlations for vital statistics—well-being, population, income, life expectancy, and rate of firearm deaths—for the 50 states in the United States.

Note that because the correlation coefficient is symmetric (e.g., the correlation between healthy life expectancy and adult mortality is the same as the correlation between adult mortality and healthy life expectancy), the correlation matrix is also symmetric across the diagonal (more on this topic in Chapter 2). Therefore, only the upper-triangular half or lower-triangular half of the matrix need to be shown. In the examples in this chapter, we highlight several different common styles of presenting correlation matrices in scholarly literature. For example, in Table 1.3, we used only the lower-triangular half of the table to show the entire correlation matrix. In Table 1.4, we showed both triangles of the table, and it can be easily verified that the correlations are symmetric by comparing equivalent correlations (e.g., compare the correlation between the first and second variable to the correlation between the second and first; both correlations equal .050).

Table 1.3 Correlations Between Income Inequality, Country-Averaged Tendency to Trust Others, and Measures of Public Health

Variable	*Income Inequality*	*Trust*	*Public Health Expenditures*	*Healthy Life Expectancy*	*Adult Mortality*
Income inequality	1.00				
Trust	−.51	1.00			
Public health expenditures	−.45	.12	1.00		
Healthy life expectancy	−.74	.48	.34	1.00	
Adult mortality	.55	−.47	−.13	−.92	1.00
Mean	0.363	3.9	5.6	68.0	0.129
SD	0.769	1.1	2.1	6.2	0.090

Note: Data came from the International Social Survey Program and included 48,641 respondents across 33 countries. Income inequality was assessed using data from the World Bank World Development Indicators database. Trust was measured as a country-level average of participants' rating of the statement "There are only a few people that I can trust completely," defined on a 5-point Likert-type scale (1 = *strongly agree*, 5 = *strongly disagree*). Public health expenditures, healthy life expectancy, and adult mortality measures were accessed using data from the World Health Organization Statistical Information System. Adapted from Elgar (2010).

Table 1.4 State-Level Vital Statistics for 2016–2017

Variable	*1*	*2*	*3*	*4*	*5*
1. Well-being	1.00	.050	.422	.708	−.416
2. Population	.050	1.00	.129	.246	−.254
3. Per capita income	.422	.129	1.00	.748	−.682
4. Life expectancy	.708	.246	.748	1.00	−.803
5. Firearm death rate	−.416	−.254	−.682	−.803	1.00
Mean	61.5	645	30.5	78.2	13.6
SD	1.20	720	4.24	1.76	5.32

Note: Well-being was measured using the Gallup-Sharecare Well-Being Index. Population was measured in units of 10,000 per estimates from the U.S. Census Bureau in 2016. Per capita income in units of $1,000 was reported for 2017 per the *Chronicle of Higher Education*. Life expectancy is reported for 2017 by National Geographic. Firearm death rate was reported for 2017 per the Centers for Disease Control and Prevention.

In each of Tables 1.3 and 1.4, we included rows showing the mean and standard deviation for each variable. We provided the additional information in these tables—and also in several future tables—so that readers can transform the correlation matrix into a covariance matrix. The formulas to transform a correlation matrix into a covariance matrix (and back again) rely on matrix algebra and are beyond the scope of treatment in the current book. We show, however, the relationship between a single correlation and its equivalent covariance:

$$\mathrm{Cov}(X, Y) = r_{XY} * SD_X * SD_Y \tag{1.4}$$

This transformation requires the correlation and the standard deviations of the two variables. To provide a computational example, consider the correlation between Well-Being (WB) and Population (POP) in Table 1.4, $r = .050$. Using Equation 1.4, the covariance can be computed to be $\mathrm{Cov}_{\mathrm{WB,\ POP}} = .050*1.20*720 = 43.2$. The covariance between Well-Being and Per Capita Income (IN) is $\mathrm{Cov}_{\mathrm{WB,\ IN}} = .422*1.20*4.24 = 2.15$. Just as unstandardized variables cannot be compared with one another (see the height–IQ example above), covariances also cannot be compared. But because covariances transformed to correlations adjust out measurement scale differences and thus become comparable, correlations can be compared with one another in a meaningful way. Thus, we can report that the relationship between Well-Being and Population ($r = .050$) is substantially weaker than the relationship between Well-Being and Per Capita Income ($r = .422$).

The two covariances (43.2 and 2.15), on the other hand, reflect both relationship differences and differences between the measurement scales and, thus, are not naturally comparable.

Consider also the correlations presented in Table 1.5, adapted from Leonard (1997), which summarize how the racial composition of a sports team is associated with the racial composition of the city that team represents for professional basketball (National Basketball Association [NBA]), football (National Football League [NFL]), and baseball (Major League Baseball [MLB]). The table is organized such that three correlation matrices (one each for basketball, football, and baseball) are collated side by side and presented with only the upper-triangular half of the matrix. Organization of the correlation coefficients into matrices, and then concisely displaying these correlation matrices simultaneously, facilitates identification of patterns in the data. For instance, although there are near-perfect correlations between the percentages of Black residents in a franchise city in 1980 and 1990 for all three sports, indicating consistency in percent minority between 1980 and 1990 for the cities in the sample, the correlations are different across the sports for the number of Black teammates on teams between 1983 and 1989. We also see that professional baseball demonstrates consistently lower correlations between racial composition of a team and racial composition of the franchise city than do professional basketball and football. Is there evidence that the associations between racial composition of cities and teams in baseball act differently from those in basketball and football in this time period? Is there evidence that the associations between racial composition of cities and teams in basketball act similarly to football? These hypotheses are most efficiently demonstrated and tested in the context of correlation matrices, rather than through the tedious and inefficient inspection of individual correlations.

The statistical methods for analyzing correlation matrices are useful for exploring how—and sometimes why—variables are intercorrelated. For example, consider the correlation matrix presented in Table 1.6 on childbearing intentions and outcomes (where we present the whole symmetric correlation matrix). The correlation matrix is slightly modified from data collected from the 1979 National Longitudinal Survey of Youth (NLSY79), a nationally representative sample that was first assessed in 1979 when youth were 14 to 22 years old; the sample has been followed at least biennially thereafter. A researcher may be interested in the roles of childbearing intentions and previous childbearing outcomes in predicting future childbearing outcomes. Methods like path analysis, or more generally SEM, are designed to investigate these underlying processes.

Table 1.5 Correlation Matrices Indicating the Association Between the Number of Black Members of a Professional Sports Team (No. Black) for a City in a Given Year and the Percentage of Black Residents of That City (% Black) in a Given Year by Professional Sport

Variable	*No. Black 1983*			*No. Black 1989*			*% Black 1980*			*% Black 1990*		
	NBA	*NFL*	*MLB*	*NBA*	*NFL*	*MLB*	*NBA*	*NFL*	*MLB*	*NBA*	*NFL*	*MLB*
No. Black 1983	1.00	1.00	1.00	.41	.13	.09	–.06	–.05	–.18	–.06	–.10	–.23
No. Black 1989				1.00	1.00	1.00	.37	.36	.11	.29	.30	.04
% Black 1980							1.00	1.00	1.00	.99	.99	.96
% Black 1990										1.00	1.00	1.00
Mean	6.26	13.2	6.69	8.04	17.8	4.81	26.8	28.6	30.6	27.5	30.7	31.9
SD	1.79	2.90	2.78	1.89	4.23	2.02	20.5	20.1	17.7	21.3	19.4	18.6

Note: Sample sizes differed by year and professional sport; for all examples in later chapters using these data, we use a sample size of $N = 26$ teams. NBA = National Basketball Association; NFL = National Football League; MLB = Major League Baseball.

Other statistical methods for analyzing correlation matrices, such as factor analysis and PCA, can be used to construct novel measures and assessments. For example, the correlation matrix (presented in its lower-triangular form) in Table 1.7 shows nine questionnaire items measured on 6,007 individuals from the NLSY79. Four of the items appear to assess positive self-esteem, three items appear to measure risk taking that may result in positive change, and the remaining two items appear to measure openness to others and new experiences. In Chapter 4, we will discuss how statistical methods can be used to construct scales to more completely explore if these items measure what we presume they measure.

Correlation matrices not only are useful for testing novel hypotheses but also can be vital to exploring patterns that might not be detected in the raw data, especially when the number of variables is exceedingly large. For example, consider the correlation matrix of index components for Standard & Poor's 500 Index, which captures the performance of 500 leading U.S. businesses and serves as a metric for how U.S. stocks are performing. Variables from these 500 companies may be difficult to summarize, let alone envision, in their raw form. Approaches such as factor analysis, PCA, and graphical methods (all of which we develop in future chapters) can help shed light on the complex associations among very large correlation matrices of this type.

Table 1.6 Correlations Between Childbearing Intentions and Childbearing Outcomes for 7,000 NLSY79 Respondents

Variable	*1*	*2*	*3*	*4*	*5*
1. Ideal number of children (1979)	1.000	.876	–.020	.484	.114
2. Expected number of children (1979)	.876	1.000	–.487	.389	.023
3. Number of children (1980)	–.020	–.487	1.000	.120	.441
4. Ideal number of children (1982)	.484	.389	.120	1.000	.207
5. Number of children (2004)	.114	.023	.441	.207	1.000
Mean	2.53	2.36	0.143	2.40	1.98
SD	1.53	1.46	0.45	1.36	1.46

Note: Ideal number of children was truncated at 5+ children. Expected number of children was truncated at 4+ children. Total number of children born to the respondent was reported in 1980 and 2004. Polychoric correlations were calculated on each variable, and missing data were pairwise deleted. NLSY79 = 1979 National Longitudinal Survey of Youth.

Table 1.7 Correlations Between Survey Items in the NLSY79 (N = 6,007)

Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1. I am a person of worth.	1.0								
2. I have a number of good qualities.	.73	1.0							
3. I have a positive attitude with myself and others.	.53	.57	1.0						
4. I am satisfied with myself.	.45	.47	.63	1.0					
5. Willing to take risks in occupation	.06	.07	.06	.04	1.0				
6. Willing to take risks in other people	.05	.05	.03	.03	.37	1.0			
7. Willing to take risks in making life changes	.03	.05	.03	.01	.50	.39	1.0		
8. Extraverted, enthusiastic	.12	.11	.17	.13	.05	.05	.08	1.0	
9. Open to new experiences, complex	.10	.10	.13	.11	.08	.03	.10	.29	1.0
Mean	1.5	1.4	1.6	1.8	3.9	4.1	4.2	5.0	5.2
SD	0.7	0.6	0.7	0.7	3.2	2.9	2.9	1.8	1.6

Note: Items 1 to 4 were measured from 1 (*strongly agree*) to 4 (*strongly disagree*) in 2006 and reverse coded for ease of interpretation. Items 5 to 7 were measured from 0 (*unwilling to take any risks*) to 10 (*fully prepared to take risks*) in 2010. Items 8 and 9 were measured from 1 (*strongly disagree*) to 7 (*strongly agree*) in 2014. NLSY79 = 1979 National Longitudinal Survey of Youth.

Summary

This book is about correlation matrices. We focus on helping the reader develop an appreciation and intuition for using correlation matrices. We are writing for the student or researcher in the social and behavioral sciences, but the contents will appeal to students and researchers in any discipline that uses correlations, and of course to statisticians and applied mathematicians as well. We do not assume advanced mathematical knowledge; an introductory graduate (or even undergraduate) course in statistics will suffice to get started with the material. We avoid references to advanced mathematical discourse, except in the few cases we believe advanced mathematics are necessary to understand the material we are presenting (and in those treatments, we are careful to warn mathematically less sophisticated students that the material may be skipped or scanned without loss of continuity). Researchers interested in deeper treatment of matrices in general are referred to a linear or matrix algebra textbook. Also outside of the scope of this particular book is deep treatment of statistical methods to analyze correlation matrices. Whole courses and many introductory and advanced textbooks are devoted to these methods, such as factor analysis (e.g., Finch, 2019; Kim & Mueller, 1978a, 1978b); PCA (Dunteman, 1989); SEM (Long, 1983; Preacher et al., 2008); and meta-analysis (Wolf, 1986). Each of these methods involves fitting models to correlation (or covariance) matrices, and we briefly review those at a conceptual level; we also refer readers who wish for more advanced treatment to appropriate references, such as those mentioned above, which are all available in the Sage Quantitative Applications in the Social Sciences (QASS) series.

The organization of this book is as follows. In Chapter 2, we explore mathematical properties of correlation matrices. We minimize throughout, the use of equations or sophisticated mathematical operations (e.g., advanced matrix algebra). This chapter is pivotal for understanding the structure and function of correlation matrices. In Chapter 3, we provide details on common null hypothesis significance tests for correlation matrices, including how to conduct these tests. In Chapter 4, we overview methods that use correlation matrices as the raw data, including factor analysis, SEM, and meta-analysis. In Chapter 5, we demonstrate graphical methods for displaying correlation matrices of varying sizes and structures, usually with reference to the correlation matrices that have been presented as examples in the current introductory chapter. In Chapter 6, we describe work on the geometric underpinnings of correlation matrices, which is where most of the recent modern study of correlation matrices in the statistical literature has been focused. This chapter can be skimmed or skipped by introductory students. Finally, in Chapter 7, we provide a short conclusion and summary of the book.