1

The Psychometric Quest

CHAPTER MENU

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
 Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

A friend and colleague of ours is a practicing psychometrician at a company specializing in leadership measurement. A few years ago, in preparation for a report in elementary school, his son asked him about his occupation. He told him he was a "psychometrician." He practiced the pronunciation of the word with him and told him the kinds of things he did. He explained that his life's work was focused on finding ways to measure important but subtle things, that is, things that are difficult to measure. That evening he asked his son how the report went. It became clear that even after practicing the word "psychometrician," he could not recall it. The best the boy could muster was to report to the class that his father was "a psycho," and that he tries to find ways to measure things that are hard to measure. His report to the family generated a good laugh with no threat to his father's professional confidence.

However, in the 1930s, there was a major threat to the integrity of psychometrics as a discipline generally, and indeed to all the behavioral and social sciences, as described in Section §6. The story is there related of the Ferguson Committee of the British Association for the Advancement of Science (physics vs. psychology) in their ongoing debate regarding the status of measurement in the social and behavioral sciences. The debate continued for years, throughout most of the 1930s. The vote of half of the committee was that it is not possible in principle to measure psychological attributes scientifically. There is much at stake here, for it calls into question whether any one of the behavioral sciences could truly be considered to be a science. Fortunately, several psychometric and statistical scholars since that time (Luce & Tukey, 1964; Luce, Krantz, Suppes, & Tversky, 1990; Krantz, Luce, Suppes, & Tversky, 1971; Suppes, Krantz, Luce, & Tversky, 1989) have successfully refuted that challenge, with assistance from recent advances in metrology (Stock, Davis, de Mirandes, & Milton, 2019). There is also the pragmatic argument. Psychometric theory and methods have, over the past century, developed into a necessary and useful branch of applied science in a variety of areas including psychological, educational, and medical testing.

In 2012, Ian Stewart (2012) published a book about "seventeen equations that changed the world." We would like to propose an eighteenth equation—the analytical method known as factoring. Without factoring, there would be no psychometrics as we know it. Factoring is the foundational method for the development of psychometric theory and practice. It began early in the 20th century with the publication of Karl Pearson's (1901) paper that introduced what later came to be called the "principal component" analytical method. A few years later, Charles Spearman's (1904b) paper introduced the parallel method that would become "factor analysis," an expanded application of the mathematics of principal component analysis to address psychometric questions. Spearman's purpose in developing the factor analytic method was to examine the elemental structure of human intelligence.

In the intervening century, factor analysis has become the essential method for examining not just intelligence but many other constructs that cannot be directly observed. It enables psychometricians to examine whether test items reliably cluster together reflecting a common underlying psychological construct. For example, do such things as knowledge of state capitals, or of vocabulary words, and the rapidity with which one can solve a puzzle inter-correlate in a way that reflects an underlying factor that could be called intelligence? Does intelligence consist of one overarching component that best describes what it means to be cognitively capable, or are there a variety of types of intelligence such as judicial wisdom, creative and mechanical skill, emotional intelligence, social facility, and interpersonal effectiveness? Perhaps intelligence is better understood as a constellation

of related but distinct neurocognitive parts such as working memory, processing speed, and verbal reasoning.

What about personality tests? Are they capable of predicting how people will respond in situations? Can they be used to guide therapeutic interventions, or to clarify a clinical diagnosis? Factor analysis has proven to be effective in addressing psychometric questions such as these in a wide variety of areas. Without factor analysis, the past century of development in psychometric theory and methods would not have been possible.

1.1 Quantification in Psychometrics: Historical Origins

1.1.1 (§1) Why Are Mathematical Methods Central to the Empirical Sciences? Why Are Statistical Methods Used in the Behavioral Sciences?

This book represents the culmination of an academic quest of nearly 50 years by the primary author, Dr. Petr Blahuš, brought to completion after his death with assistance from several of his friends. Central to his quest has been a belief in the essential unity of all scientific endeavor. In his own words:

> I should introduce myself as an advocate of the belief that there exists a general unity of scientific knowledge. This monistic personal philosophy of mine stresses that *in principle*, there exists an *essential unity of scientific method* in the empirical sciences. This philosophical view has nothing to do with positivism, and especially not with its two extreme versions: *scientism*, which claims that there is an absolute identity in the methods used, and *reductionism*, which moreover states that the identity is also found in the object under study (i.e. psychology can be finally reduced to the principles of physics). Therefore, my personal philosophy could be perhaps called a relative and moderate methodological monism.

Petr maintained that this essential unity permits us, step by step, to walk on the endless path of learning, to improve our limited knowledge and to find connections among the many and varied branches of science. The pieces of our knowledge, though insufficient and temporary, must be organized or *systemized* so that we can find and recognize certain regularities. In this systemizing effort, certain auxiliary devices are not only helpful but necessary. These auxiliary devices are the *formal methods* of the abstract sciences, including mathematics, logic, and a variety of graphical devices such as the diagrams of molecular structure. These formal methods are essential to the pursuit of empirical science. Complex and subtle relationships among empirical observations cannot be uncovered and described without them. Imagine trying to do chemistry without structural diagrams of molecules or atoms. Or try to imagine a world in which the science of physics must be pursued without the availability of calculus.

The myriad branches of the empirical sciences are found at various stages of progress in their use of these formal methods. In past centuries, most applications of formal methods came primarily from quantitative and deterministic mathematics. They were also closely related to specific empirical problems in the established fundamental sciences, such as classical physics and astronomy. Later, and even up to the present, methods adapted to qualitative concepts and able to deal with a certain degree of uncertainty were employed in the human sciences such as psychology. Because these methods are not so closely related to specific research applications, they have broader and more flexible applicability. Statistical methods have become the prototype. They are typically the beginning of a formalistic infiltration into a branch of the human empirical sciences, that is, behavioral or social sciences such as psychology, kinesiology, sociology, political science, economics, organizational behavior, and education.

1.1.2 (§2) The Specific Mission of Statistical Methods in the Behavioral Sciences

The role of statistics, as with all kinds of formal methods, must be determined with regard to the overall purpose of the particular science. In short, the main purpose of all sciences is to approach an *understanding* of nature and to disclose the *laws* by which nature is governed, as far as possible due to the partiality which is allowed to us. In the empirical sciences, that purpose is sometimes pursued by a general methodology that includes the two fundamental scientific functions of *explanation* and *prediction*. The explanation is typically based on *general scientific laws*, or even *universal laws*, which might even have a *causal* character deeply rooted in the nature of the observed phenomena. On the other hand, prediction[1] often may be based on certain superficial regularities. These *partial and descriptive superficial empirical regularities* can offer empirical estimates, quite often directly for some practical use and possibly without any deeper explanation.

As we can see, all these procedures are based upon a formalized *systemization*. Thus, we could characterize the role *of statistical methods* in the social and behavioral sciences as that of *contributing to our knowledge of scientific laws and regularities* through formal symbolization. This is carried out on two levels. One is the *exploratory* role in discovering new regularities. The second one is *confirmatory*, that is, testing the theoretically derived hypotheses with a special aim not just to confirm or reject but to specify a *form of the lawfulness* under scope. Here the "form" should not be understood too narrowly, say as a curve relating two quantitative variables to one another. The specification of form can be accomplished in a wide variety of ways, including verbal description, graphics and diagrams, and mathematical expressions.

Design is a complementary and necessary third component to the main two functions, that is, to the *exploration* that leads to hypotheses, and to the *confirmation or disconfirmation* of proposed explanations. To obtain a confirmation of a hypothesized theoretical proposition, a meaningful explanation, requires both careful instrument design and careful experimental design to set up the confirmation situation and rule out alternative explanations. Thus, we might say the processes of *exploration* and *explanation* require the *design* to attain reproducible scientific merit.

In this way, statistical methods, as well as other formal methods, have become part of a gargantuan network of many approaches to systemizing the relationships and links between observable data and scientific theory. Gradually, the formal apparatus and the original empirical core of the branches of science are becoming amalgamated and mutually synthesized. In this process, the separate and occasional applications are transformed into several foci of true interdisciplinary scientific disciplines. In these focal synthesized areas, the "empirical-formal tandem" works in cooperation that allows each of the components to influence the other and to initiate mutual development in a creative way.

1.1.3 (§3) Why Do We Use *Behavioral Statistics* and Not Just a Branch of Applied Mathematics?

This is a reasonable question, for "statistics" is being used as an abbreviation for "mathematical statistics," which didn't become a fully recognized mathematical discipline with axiomatic background until the end of the 19th century and the beginning of the 20th century. One aspect of the answer has been indicated in the last section: some branches of science require an adjustment or

1 We could use the term prediction in a broader sense, almost in a metaphorical sense of "the guessing of the unknown." This may give a connotation of "forecasting" that would be misleading. A better term would be *estimation*, which would cover all three time-horizons: (i) *contemporary estimation* as of diagnosing a psychotic syndrome from certain symptoms, (ii) prediction in its narrower sense of forecasting future success in school from admission tests, or (iii) retrodiction of a certain prenatal brain injury based upon symptoms found in adulthood.

adaptation in the mathematical procedures to accommodate an uncertainty component in prediction and toleration for a broader variety of approaches to quantification. Also, there are several relatively stable synthesized interdisciplinary areas that psychology and statistical science have in common. You find them, for instance, in psychophysics in the form of theories and models for the measurement of sensory stimuli. They also are found in psychometric methods such as classical test theory (CTT) models and item response theory models, or latent structure models of intelligence.

Another aspect of the fit between the behavioral sciences and the discipline of statistics can be illustrated in the historical development and mutual influence of the two disciplines. Their relationship is really unique and could be compared to the historical relationship between classical physics and deterministic mathematics including geometry. In some social and behavioral sciences, the use of statistics started in terms of isolated applications which helped further the development of empirical science. However, with psychology, the situation was quite different. Historically, psychologists themselves tried to create statistical methods for solving their specific research problems and this became an impetus for the consequent mathematical treatment of the problem on a more formally sophisticated basis. This can be illustrated with the example of the mutual interconnection between methods of correlation and theories of intelligence. Without the pressure and demand that psychological questions imposed upon statistics and mathematics, the development of methods such as correlation, regression analysis, factor analysis, and related latent-variable developments would have certainly been delayed, if not precluded.

Furthermore, the synthesis of psychology with statistics is not limited to meeting the general aims of scientific inquiry. Psychology also has very practical clinical branches, which make it parallel to similar situations in medicine with regard to biological science. In these areas, the intention is to produce practical benefits through clinical treatment. But every treatment must be preceded by diagnosis, and for that reason, psychometrics has become an essential tool in many kinds of clinical work. A neuropsychologist who specializes in the diagnostic evaluation of nervous system disorders is a good example of this. In a sense, "psychometrics" is an interdisciplinary theory of behavioral diagnostics, a synthesis of psychology, psychobiology, psychiatric medicine, and statistics. From another, and perhaps an oversimplified, point of view, psychometrics could be described as the theory of measurement errors in behavioral evaluation technology. The psychometric theory of psychological tests and questionnaire inventories, dealing as it does with the problems of validity, reliability, and measurement error, can be introduced as a model illustration of the genuine interdisciplinary synthesis of an empirical "soft" social science with the formal "hard" sciences of quantification.

1.1.4 (§4) Some Remarks on the Historical Relationships Between Psychology and Statistics: The Development of the Correlation Coefficient and Least Squares Regression

The earliest and most fundamental contributions to the science of psychometrics date to the second half and end of the 19th century, and the beginning of the 20th century. Under the influence of Darwin's theory of evolution, the mutual relationship between the influence of heredity and environment began to be studied. There were two main lines of study and two great men who accomplished major advances in genetics in the last half of the 19th century. The first was Gregor Mendel, a monk who recorded the effects of hybridization on qualitative attributes (like color) of beans and peas which he grew in the monastery garden in Brno (a Czech town in the Bohemian Kingdom, in those times a part of the Austrian Empire). More important for psychology, however, was the second man—an Englishman by the name of Francis Galton—whose prodigious and varied scientific

labors were partially enabled by being born into wealth. His early accomplishments were in geography, but, influenced by his half-cousin Charles Darwin, his attention turned to human biology, anthropology, and intelligence, with a particular interest in the problem of "genius" and its hereditary roots (Galton 1869, 1889). In contrast to Mendel's qualitative approach, Galton focused on quantitative, measurable attributes and used them to compare inter-individual differences among consecutive human generations, with special attention to individual "deviations" from the population mean. Many years later, the famed statistician Sir Ronald A. Fisher explained the mathematical inter-relationship between the Galtonian and Mendelian approaches, in a classic paper containing the first use of the term "variance" (Fisher, 1918).

1.1.4.1 The Correlation Coefficient and Least Squares Regression

Influenced by phrenological assumptions about the relationship between abilities and the shape of the skull, Galton attempted to evaluate the degree of heredity by the similarity of skulls between fathers and sons. As well, he invented tests to measure several psychological abilities, especially what he called "natural (mental) ability," a predecessor of IQ, which, he believed, was primarily innate. He tried to find mutual cross-generational similarities as indicated by corresponding father and son deviations above or below their generation's average. For this purpose, he invented a quantitative index of statistical similarity. Later, with mathematician, philosopher, and his close junior friend, Karl Pearson, they jointly improved the index to be between 0 and 1 in absolute magnitude and called it the coefficient of "co-relation" (Galton, 1888).

This index is well known today as the Pearson product-moment *correlation coefficient* (Pearson, 1896), or the PPMCC (see Appendix 2). The name was derived from the formal similarity of its computational formula to the mathematical formulas for mechanical moments[2] in classical physics. Mathematically, the correlation coefficient is based upon the "least squares" method.

Much earlier, in the late 18th and early 19th centuries, two other men formulated and applied the least squares mathematical method. It would later become least squares regression analysis (see Appendix 4), which in its current form is closely related to the correlation coefficient. The first of these two men was Carl Gauss, who in his astronomical observations of the orbits of planets in 1794 or 1795 worked on the problem of *errors of observation*. He tried to fit and smooth the imperfectly observed measurements in a mathematically optimal way. Thus, while the exact *true* orbits remained unknown, they were best approximated by the fitted values, with optimal fitting achieved by minimizing the sum of squared deviations of the observed values from the fitted values. The second man was Adrien-Marie Legendre (1805), who published the first paper on the topic.[3] Although the discovery of this "least squares" method preceded the development of the correlation coefficient by nearly a century, in modern treatments of the two topics, the correlation coefficient is typically introduced first, and then the computational method for least squares regression is taught as an extension of the logic and mathematics of the correlation coefficient, as shown by Appendix 4 (regression) being preceded by Appendix 2 (correlation) in this book.

2 Regarding the correlation coefficient as an application of mechanical moment calculations from physics Petr Blahuš, our senior author quipped, "So, do not worry if some time you meet a strict statistician who tends to stay on the ground of pure mathematics as far as possible and explains the correlation coefficient as 'simply a central standardized second-order mixed moment,' and then adds, 'perhaps you could use it somehow in psychological research, too'."

3 Gauss is credited by many with being the discoverer of the least-squares method. However, some would argue that Legendre should be accorded that honor, since he published the first paper on the topic. See Plackett (1972) for a weighing of the arguments and evidence in favor of each side. See also Stigler (1981) for additional information that has come to light, and Brown, Hendrix, Hedges, and Smith (2012, p. 373) for a two-paragraph summary of this major priority dispute in the history of statistics.

1.1.4.2 Correlation, Regression, and Heritability

The explicit formulation of the correlation coefficient expressing the degree of association between various pairs of human attributes, *x* and *y*, was motivated and pursued by Galton, and then mathematically formulated by Pearson. The degree of association between fathers and sons was interpreted as the amount of their common genetic background—the heritability—with regard to a certain attribute. The next step was to focus on the problem of which attributes are more heritable and which less. Intelligence, which he referred to as "mental ability" should have been mostly innate, according to Galton's theory, perhaps more than other attributes. Therefore, it was of particular interest to compare the correlation, that is, the degree of heritability, in this attribute. Galton devised the first scientific mental measurement tests, making him the founder of psychometrics. He also created a wide variety of "anthropometric" measurements of the chief characteristics of the human body, which could also be tested for heritability in comparison with the heritability of mental characteristics.

Thousands of subjects were tested with Galton's anthropometric and mental measurement methods at health exhibitions, and in his laboratory at the South Kensington Museum. Many were willing to pay a small fee for the privilege of participating and receiving a copy of their results. He gathered so much data that it was not analyzed until well into the 20th century. Galton's methods were somewhat consonant with the phrenological speculations of the times, that is, the idea that the form and size of the skull had a direct relationship to "mental faculties." For example, a higher father-son correlation between skull circumferences than between their body heights might support the hypothesis of a higher heritability of "mental ability" than of physical body characteristics. The Gaussian least-squares fitted line enabled the estimation or prediction of a son's head circumference from that of his father.

Galton's interest was in developing methods for the prediction of mental capabilities from one generation to the next. The relationship was thought to be decomposable into a *predicted* part, which he hoped would be due to *heredity*, and an *unpredictable* part, which he presumed would be due to the *environment*. This was an *important step in the general methodology of science*. It was the beginning of the turn from a *deterministic approach* to measurement, as in physical science, which had allowed only the researcher's errors of observation as a kind of methodological (epistemological) imperfection of science, to a new incorporation of the *stochastic*[4] *conception of the world*, allowing some *ontological uncertainty* (that is, actual uncertainty in the makeup of the laws of the universe) to be an essential component of nature and its laws.

Of course, the controversy of epistemological vs. ontological uncertainty was well known as a philosophical problem well before the time of Galton. In some well-established sciences, such as physics, it was discussed much earlier. For instance, Laplace (1814/1952) articulated the view that uncertainty is at the root epistemological. He proposed that if we were able to "know all the forces that set nature in motion and all positions of all items of which nature is composed" then we would have the future before our eyes. On the other side, and more recently, Peirce (1901/1957) articulated the ontological view. He proposed that uncertainty is at least partly inherent in Mother Nature—a property of natural law, including the laws of physics. Galton's achievement was to bring a stochastic emphasis into the life sciences and the behavioral sciences. In this, he stood on the same side as Peirce.

The fitted estimation line, which was intended to make a graphical estimate from fathers' to sons' values of head circumferences, was called by Galton a "reversion line." The estimate could also be produced arithmetically by a simple equation that describes the line. The *estimation equation* as a

4 A process that is stochastic could also be referred to as "probabilistic," randomly determined, or at least somewhat attributable to random chance as opposed to determinism.

device for producing the numerical prediction of sons' from fathers' head measurement values is shown here using fictional illustrative data, numerically assumed to be in inches.

$$(\textit{predicted}\ \text{son's head circumference}) = 1.02\,(\text{father's head circumference}) + 0.04$$

$$\hat{Y} = 1.02X + 0.04$$

In general, the equation for (linear) statistical estimation contains two coefficients a, and b, with the resulting estimated value y marked by a hat ^, which signifies that it is an estimated value.

$$\hat{Y} = bX + a$$

The numerical values of the above coefficients can be interpreted as follows:

$a = 0.04$ means that there is a tendency for the generation of sons to have a slightly larger head circumference than the fathers, a 0.04-inch shift,

$b = 1.02$ tells us that a father who has a 1-inch larger skull than another father should have a son whose skull circumference is, on average, about 1.02 inches larger than the son of the other father.

Moreover, coefficient b also means that the inter-individual spread among sons is larger than the inter-individual spread in the fathers' generation.

Galton's observation that the predicted values $\hat{y}$ for the sons of fathers with particularly high or low deviations from the group mean tended to be closer to the mean of the new generation was called "regression toward the mean."[5] He interpreted this phenomenon as an effect of, and evidence for, Darwin's theory regarding the influence of natural selection. The name stuck. In modern statistical practice, the best-fitting linear estimation equation is now referred to as the *regression equation*, and its geometric representation is the *regression line*.

In his 14 June 1920 lecture to the Royal Institution, Pearson himself informs us that the symbol r that we now use for the Pearson product moment correlation coefficient also comes to us from Galton: "Here for the first time appears a numerical measure r of what is termed 'reversion' and which Galton later termed 'regression'." This r is the source of our symbol for the correlation coefficient, which was really the letter of "reversion" not of "regression" (Pearson, 1920, p. 33; see also Piovani, 2008). In this same paper, Pearson corrects his earlier error in crediting Bravais (1846) and his product sum with the discovery of the correlation coefficient and accords that honor to Galton.

The concept of correlation, as well as the Pearson product-moment correlation coefficient as a numerical index, is now essential and ubiquitous. It has become indispensable in practice as a data-analytic tool, but it is also central and vital as a theoretical concept. As the absolute value of r approaches 1, the accuracy of the regression estimation grows to its maximum. If the accuracy reaches a sufficient level, the equation can be used for *practical estimation*, that is, for prediction. This descriptive statistic, r, enables us to evaluate the predictive validity of psychological diagnostic methods, thus providing a rational foundation for using a test to predict the "fit" of an applicant for a certain job or perhaps using a test to predict how well a student will do in graduate school.

Correlation and regression are the fundamental underlying concepts of the modern psychometric theory of tests. They make it possible to establish the reliability and the possible diagnostic error of a newly constructed test battery or questionnaire. They are also central to a host of other diagnostic quantities such as dimensionality, construct validity, and consistency. At a more theoretical level,

5 This statistical phenomenon, oft-observed, can be stated simply: if an observation of a random variable is extreme on first measurement, it will generally be closer to the mean on the next measurement.

they constitute the building blocks for the wide variety of factor analytic methods, as well as the path-analytic structural equation models and the associated latent variables that enable one to identify theoretical scientific concepts, referred to as *constructs*.

1.2 Variables, Scales, and Data

In the terminology of behavioral research methods, there are several areas that are affected by semantic ambiguity and confusion. The problem has deep methodological roots, and creates several areas of confusion, with important and troublesome consequences. We will clarify several of the terminological issues in Section §5 that follows.

Following the introduction of these clarification issues, Sections §6 and §7 give a brief historical account and discussion of a central question for psychometrics, "Is measurement in the social and behavioral sciences even possible?" and the corollary question of whether these disciplines, if they do not have bona fide measurement, could appropriately be considered sciences.

1.2.1 (§5) *On Terminology I*: Attributes with Categories, Variables with Values, Tests with Scores, and What About "Data"?

The term "score" is usually used narrowly to refer to the results of "tests" in the sense of psychological tests, such as sensorimotor tests, admission tests, standardized questionnaires, and so forth. However, in the behavioral, educational, and social sciences generally, *Test Theory* in its broad meaning is conceptualized as a *general theo*ry of the *diagnostic quality* and *standardization* of *any* method used to measure or classify behavior, whether in empirical research or in clinical or educational practice. Besides the term "scores" there are also other related terms, such as "values," "categories," "alternatives," or "levels," used to refer to types of behavioral classifications or measures.

1.2.1.1 Variables and Data

Another ambiguously used term is "variable," which takes on a wide range of somewhat differing meanings. For some people "a variable" is still joined with the idea of a mathematical variable, taking on many possible values, and being used in contrast or opposition to a mathematical "constant" which has a single value. Such a conception, which focuses on quantity, and distinguishes between a variable *quantity* and a constant *quantity*, is too narrow for a general methodology, due to its automatic assumption that we are dealing only with the well-defined numerical quantities of physics and other foundational empirical sciences. Moreover, it suggests the naïve and simplistic idea of the metric axis represented by real numbers, which is taught in elementary mathematics in school.

Is gender to be considered a variable? Is ethnicity a variable? Each obviously must be considered as such and have a place in the social or behavioral sciences if these sciences are to be based upon empirical observation. However, even at first glance, there is an obvious problem. There is an intuitive mismatch between, on the one hand, "*gender*" as an independent *variable X* predicting "*heart attack*" as a dependent *variable Y* (with both as empirical binary categories), and those simple independent and dependent numerical/mathematical variables X and Y which everybody knows from high school mathematics. The sense in which the term "variable" is used in these two settings is obviously quite different.

Speaking in a simplistic way, the mutual relationship between those two notions of "variable" is the topic studied by a *representational* theory of measurement (RTM) (Section §6), which deals with the problem of how well numbers *represent* or reflect empirical reality, and, on the other hand,

how well a real-world attribute may be represented by numbers, that is, by the formal ideal of a numerical scale. Therefore, it is first necessary to distinguish between two worlds, namely the *substantive world* of actual things, actual *entities* and their *attributes*, which stand as an original *reality* to be explained; versus the *formal world of symbols*, very often *numbers*, with its *syntactic* rules such as arithmetic operations, through which we try to represent the first world, the world of reality.

Thus, to explicate the notion of "variable," we must distinguish between (a) the *empirical attribute* itself, for example, "strength of attitude," which is thought of in its "natural," that is nonformalized original existence; and (b) the *formalized attribute,* as coded by symbols such as numbers. The primitive empirical attribute (a) could be provisionally organized by dividing it into *categories*. In this conception, we define the *category* as a *discrete possible alternative* outcome that an *unformalized attribute* may take on. However, in this sense, a category should *not* be assumed to be only purely qualitative (as in some statistical texts, where the purely qualitative "categorical data" are treated in a special way). Thus, even an attribute of a fully quantitative and continuous nature, like body weight, could be assumed and used as discrete. It could be used to sort people into categories given by some level of resolution. A subject's body weight numerically formalized as 140 lb. could mean it falls into the category represented by an interval between 139.5 and 140.5.

The formalized attributes (b), usually consist of symbols that can be easily processed. For example, in computerized processing perhaps ASCII characters would be preferred in general, and in particular, *numbers* are usually the most convenient. The formalization may occur as in the case of so-called keyed categories in an inventory with pre-printed response options, such as Likert scales, multiple-choice items, and so forth. Thus, the notion of a *mathematical variable quantity*—as known from high school mathematics—is just a very particular case of such a formalization. For some people, the sole act of formalization or symbolization is enough to allow the use of the term "variable" for any attribute, even if it is unclearly defined. Some are content to use terms like "values" or "data" to refer to the formal symbols, especially numbers, assigned to the categories in any ambiguous or even rather mysterious way. It is the task of measurement theory to clear up such ambiguities.

Another confusion may come from homonyms, such as when the word "*data*" is given several fundamentally *different and technically specific meanings* in different areas. In *computer processing*, for example, a programmer might refer to any information necessary for running a program as the "input data." For instance, the name of the data set, identification codes for names of subjects, etc., are also part of the input to the program. However, these are not genuine research data of scientific interest. Therefore, it is necessary to be careful in a discussion with specialists of different fields to distinguish between genuine research data and the "data," which constitute the input to a computer. Similarly, numbers on a sheet of paper might be assumed to be "the data" and automatically processed by a careless would-be statistician as if they were fully metric.

Another specific meaning of the term "data" is its use in latent variable modeling. The "data to be modeled," as input to a modeling procedure, need not be the directly observed or measured values but could be already-computed values, that is, values derived from the originally observed and coded ones. For someone skilled in latent variable modeling, for example, correlation coefficients can stand as the target "data" to be modeled or explained in terms of a latent common factors model such as one of the methods of factor analysis.

In addition, occasionally in *qualitative research* any laboratory records or field-research documentation, such as videotaped records of the behavior of children playing, or tape-recordings of narrative interviews, are sometimes called "data," in a kind of short-cut jargon. It might be wiser to refer to these kinds of preliminary observations as something like *raw research material* to avoid confusion. Some qualitative researchers advocate the definition of data as "everything that has been recorded" (Majerova & Meyer, 1999, pp. 135–141). We would suggest that perhaps the term "data"

should be reserved for the transformed state of such raw research material after it has been coded through a well-planned and logically defensible process. There is a contradictory aspect to referring to raw observations as data. Would we say that after obtaining data we then apply a coding procedure to the data? Would we then need to distinguish between the "uncoded data" and "coded data"? It seems somewhat unusual that "analyzing data" might result in their "coding." It is more usual to think of data as being ready for analysis *after* they have been coded.

What has been said about the importance of coding in qualitative research applies to every area of behavioral research. Data coding is a crucial step in the scientific process. Similar data coding, verification, and evaluation processes are also central to quantitative and even experimental research. If this foundational process is slighted, the later stages in the scientific process—data analysis, interpretation, and theoretical integration—are compromised. Haphazard or arbitrary data construction systems or strategies that are not well-reasoned may not fulfill the necessary conditions and requirements for measurement, and hence cannot be recognized and treated as "genuine data."

1.2.1.2 Scientific Rigor, Representation, and Stochastic Modeling

The question of whether various approaches to formalization may be recognized as producing "data," that is, bona fide research data that can support analysis within the framework of behavioral statistics and test theory, depends upon three broad preconditions: (i) scientific rigor, (ii) principles of representation/meaningfulness, and (iii) stochastic modeling.

The first area, scientific rigor, deals with general scientific foundations and can be considered in two parts: (ii) *standardization* of variables, and (ii) *theory-based anchoring* of variables. *Standardization* of the observational procedures includes several topics from classical measurement theory, such as establishing the reliability and validity of the variables and extending to other quality control methods that are essential to the art and science of testing. The explanation of these concepts will require additional preliminary knowledge. Hence, the discussion will resume later in the book. *Theory-based anchoring* of the variables has to do with so-called "construct validity," that is, establishing a variable in relation to other variables within the network of scientific lawfulness. It has to do with questions of scientific theory building and its corroboration by empirical data. These principles will also be dealt with in later sections, particularly in Chapter 3.

Whereas these two aspects of "scientific rigor" relate to the *practice* of psychometrics and will constitute a major part of what is taught in the remainder of the book; the remaining two preconditions—representation/meaningfulness and stochastic modeling—are both primarily *formal*. Representation and meaningfulness have to do with logical and philosophical foundations and will be dealt with in the discussion of representational theories of measurement (RTM) in Section §6.

Stochastic modeling is also formal. It is firmly based in statistical probability theory (discussed in Sections §7 and §13) and is closely associated with the concept of a random variable.[6] Traub (1997) identifies it as number two of the three major developments that made the emergence of psychometric theory possible. The other two are a recognition of the presence of errors in measurements and a concept and index of correlation. In Section §7, we will briefly explain why the decomposition of observed test scores into "true score" and "measurement error," and the recognition of measurement error as a stochastic random variable, is so important in establishing bonafide research variables. This theme will continue to appear in various places throughout the remainder of the book.

6 See Section §13, particularly the Section titled "Classical probability theory and the binomial distribution," for simple examples of stochastic modeling, which is one of the three preconditions for "bona fide research data" as introduced in this subsection of Section §5.

1.2.2 (§6) Measurement Scales and the Representational Theory of Measurement (RTM)

1.2.2.1 The Ferguson Committee

In 1932, the Ferguson Committee of the British Association for the Advancement of Science was appointed to consider and hopefully resolve the question, "Is it possible to measure sensory events, that is, human sensory experience?" Much is at stake in this question, for it is really the question of whether psychology or for that matter any of the behavioral or social sciences (sociology, political science, economics, etc.) can be considered to be sciences. The committee, consisting of delegates from physics (Section A of the association) and psychology (Section J), met for seven years with little progress. One committee member, complaining about the rigidity of opinions and the spirit of disagreement in the meetings, quipped that year after year, his colleagues "came out by that same door as they went in" (Stevens, 1946, p. 677).

In its last year of debate, the committee focused upon a particular psychophysical scale, the Sone scale of loudness (Stevens & Davis, 1938), which claimed to establish a mathematical relationship between an established physical measure of sound intensity, and a subjective scale of auditory sensations. As expected, in this final year the responses of the committee were again bifurcated and extreme, and again divided into the same two opposing camps. A statement from the negative camp is particularly illuminating, and on first look even compelling, arguing that "any law purporting to express a quantitative relation between sensation intensity and stimulus intensity is not merely false but is, in fact, meaningless unless and until a meaning can be given to the concept of addition as applied to sensation" (Final Report, p. 245).

The focus in this statement on "the concept of addition" as the central issue in whether a measurement method can be considered scientific mirrors N. R. Campbell's (1920, 1928) philosophy of measurement that held sway for over 40 years. Drawing upon the distinction between *extensive* measurement and *intensive* measurement, Campbell held that *fundamental* extensive measurement,[7] that is, acceptable scientific measurement, requires an empirical *concatenation* operation that corresponds to the numerical operation of addition, a goal beyond the reach of subjective measurement methods. The measurement of weight provides a simple example of what Campbell means by additivity in fundamental extensive measurement. Consider a beam balance with a 3-g weight in the left pan, tipped to the right by having a 5-g weight in the right pan. We then "concatenate," that is add, a 2-gram weight to the left pan and the weights are now in balance. The numerical addition operation of $3 + 2 = 5$ is *isomorphic* with the empirical concatenation operation. The two correspond to one another.

It could be said that the empirical concatenation of weights in a beam balance and the plus operation on real numbers form an isomorphism with one another. However, if this kind of concatenation operation is a *necessary* condition for establishing a valid scientific measurement scale it would follow that most psychological variables are not scientifically measurable. For example, it would be difficult to imagine what "empirical concatenation" could mean for IQ scores (are two persons with an IQ of 100 the equivalent of a person with an IQ of 200?), or for that matter, even for loudness judgments. The Section A members of the Ferguson Committee were asking the obvious next question. If those doing behavioral and social research cannot establish the

7 Fundamental measurement refers to measurement that is primary as are length and weight, not derived from other measurements as are volume and density. "Ostensive fundamental measurement" is measurement that is primary, but also concrete and demonstrable, as when for example we actually demonstrate measurement of length by laying down a yardstick repeatedly in measuring the length and width of a garden. The word "ostensive" denotes a way of defining by direct demonstration, that is, by pointing. The example of gram weights in a balance given in this paragraph is also an example of ostensive measurement.

scientific/rational basis for their *measurements* with a concatenation demonstration, how can they then use the scientific method, including measurement and experimentation, in their research?

1.2.2.2 Stevens' Representational Theory of Measurement (RTM) and of Scale Types

Stevens' response to the Ferguson Committee's evaluation of the Sone scale of loudness, and the general question of whether sensory events can be measured, appeared six years after the Ferguson report, in what was to become one of the most influential measurement papers of the 20th century. Stevens' paper, published in *Science* in 1946, was entitled "On the Theory of Scales of Measurement." The paper begins, as does this section of the chapter, with the story of the Ferguson Committee. The entire first page is focused on making connection with the issues that were of concern to the committee, and to Norman Campbell, an influential member of the committee. The Stevens paper is not, however, a head-on rebuttal of the committee's rather narrow definition of what constitutes scientific measurement. That firm corrections work would remain to be accomplished by the *axiomatic* representational measurement theorists during the last half of the 20th century, beginning with Luce and Tukey (1964) and their demonstration that additive conjoint measurement (ACM) could produce additive scales without concatenation.

Stevens' approach in 1946 was to shift the object of inquiry from concatenation operations to the identification and explication of various types of measurement scales. He begins, in the third paragraph of the paper, by agreeing with Campbell's relational theory of measurement, and then proposes ways to broaden it.

> Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale.

Clearly, Stevens' strategy here is not to directly oppose Campbell's definition of measurement as requiring concatenation, but simply to propose that it is only one of several ways to assign numerals according to rules. This is then followed up with twin questions of what other rule systems there might be, and what mathematical properties each would have. Stevens then proposes the practical aim of determining from the mathematical properties which statistical or mathematical analyses are logically admissible for each type of scale. In other words, his aim seems to be to present a few types of measurement scales, review their properties, and offer guidance regarding logically appropriate types of analysis for each. This pragmatic turn may be in large part responsible for the popularity and broad influence of Stevens' approach to RTM *scales*.

1.2.2.3 Stevens' Four Measurement Scales

The diagram in Figure 1.1 is a summary of Stevens' four measurement scales—nominal, ordinal, interval, and ratio—and their relationship to the "basic empirical operations" that characterize each. The simplest and least informative scale is the nominal scale on the left. It utilizes only one of the properties of numbers, identity (= and $\neq$), that is, whether two objects belong to the same category or not. This scale is essentially equivalent to what we refer to as "categorical" data. Stevens indicates that the nominal scale "represents the most unrestricted assignment of numerals. The numerals are used only as labels or type numbers, and words or letters would serve as well."

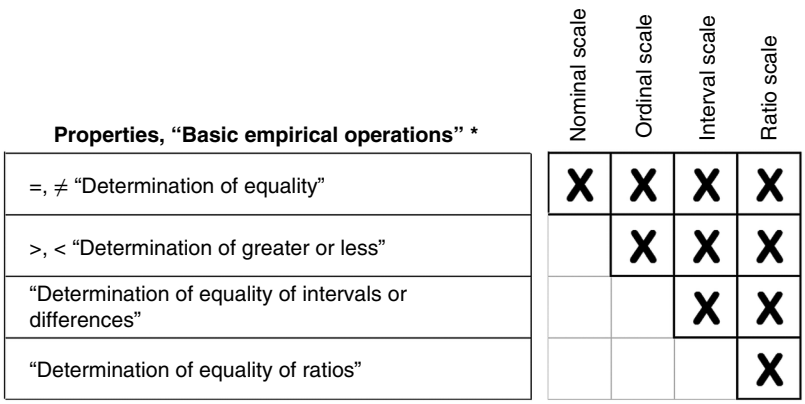


Figure 1.1 A summary of Stevens' four measurement scales.

Examples would be the numbers on jerseys in various sports such as hockey, baseball, and football, which are numerical. Non-numerical nominal scales could include professional or occupational groups such as medical doctors, attorneys, accountants, and so forth; or national groups such as Norwegians or Chinese.

Some non-numerical categories, such as college freshmen, sophomores, juniors, and seniors, are ordered categories. Senior standing is greater than junior standing, which is greater than sophomore standing, which is greater than freshman standing. These ordered categories would constitute an ordinal scale as shown in the second column of Figure 1.1, which adds a second property "determination of greater than or less than" (> and <). This is sometimes called the property of magnitude. In general, as we move from left to right (nominal to ratio) in the diagram the properties are cumulative, so an ordinal scale has the "equal/unequal" property of the nominal scale but adds to it the property of "greater or less than."

Other examples of non-numerical categories that constitute ordinal scales are Olympic medals where gold > silver > bronze or placement in a race where first place > second place > third place. Obviously, this last example of placement in a race could also be done with numbers, from 1 to whatever number is the last place, but also obviously the properties of the scale would not include additivity, since the number is only an ordinal ranking. This brings up the question of what kinds of mathematical operations make sense with ordinal scales. Stevens comments that:

> The *ordinal scale* arises from the operation of rank ordering. Since any 'order-preserving' transformation will leave the scale form invariant, this scale has the structure of what may be called the order-preserving group. A classic example of an ordinal scale is the scale of the hardness of minerals. Other instances are found among scales of intelligence, personality traits, grade or quality of leather, etc.
>
> As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety, the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply knowledge of something more than the relative rank order of data. On the other hand, for this 'illegal' statisticizing, there can be invoked a kind of pragmatic sanction: In numerous instances, it leads to fruitful results. While the outlawing of this procedure would probably serve no

> good purpose, it is proper to point out that means and standard deviations computed on an ordinal scale are in error to the extent that the successive intervals on the scale are unequal in size. When only the rank order of data is known, we should proceed cautiously with our statistics, especially the conclusions we draw from them.
>
> (Stevens, 1946, p. 679)

The major dividing line among these scale types is right in the middle of the four, between the ordinal scale and the interval scale. As Stevens says, it is at the interval level that "we come to a form that is 'quantitative' in the ordinary sense of the word. Almost all the usual statistical measures are applicable here unless they are the kinds that imply a knowledge of a 'true' zero point" (p. 679). Stevens uses the example of Fahrenheit and Celsius temperatures to illustrate the interval scale. Indeed, nearly all textbooks use this same example. That is because almost all other scales that possess interval properties also possess ratio properties. Lengths of tables are an example of a ratio scale. A table 4 m long is twice as long as one 2 m long. An interval scale would not have this property.

Temperature is a good example of an interval scale because it is so obvious that ratio comparisons cannot be made. Suppose that the high temperature in your town for yesterday was 35° and the high for today is 70°. You might be tempted to say that it is twice as warm today as yesterday, but when you convert the temperatures to Celsius, it now becomes 1.66° yesterday and 21.11° today, which would make it 12.7 times as warm today as yesterday!

In an interval scale, however, the intervals themselves, that is differences between magnitude values, *do* have meaningful ratio comparisons. For example, suppose there are two cups of water set at 78 °F. The temperature of the first cup of water is raised to 82 °F, and the temperature of the second one is raised to 80 °F. It could be said that the temperature increase for the first cup of water was twice as much as the second. The two intervals of change are in the ratio of two to one.

Narens (2002) gives a summary and critique (pp. 46–50) of Stevens' paper and appropriately entitles it "Stevens' Theory of Scales and Meaningful Statistics." He states that "Stevens, in his paper and other writings, is quite vague about what in general are the proper rules for assigning numbers to empirical objects—that is, what should be proper methods of measurement. He gives many specific examples, but this clearly is inadequate for the rigorous founding of a theory of measurement." This rigorous founding for a theory of measurement would come in another 15 years with the publication of the first volume of the *Foundations of Measurement* series (Krantz et al., 1971). Nevertheless, Stevens' place as a pioneer for broadening the approach to RTM in the human sciences is secure, clearly setting out for a wide audience how empirical relationships can be represented within, and isomorphic to, numerical structure.[8]

Scales are possible in the first place only because there is a certain isomorphism between what we can do with the aspects of objects and the properties of the numerical series. In dealing with the aspects of objects we invoke empirical operations for determining equality (classifying), for rank ordering, and for determining when differences and when ratios between the aspects of objects are equal. The conventional series of numerals yields analogous operations. We can identify the members of a numeral series and classify them. We know their order as given by convention. We can determine equal differences, as $8 - 6 = 4 - 2$, and equal ratios, as $8/4 = 6/3$.

8 Note the similarity of this paragraph from page 677 of Stevens (1946) with Michell's (1990) characterization of representational theories of measurement in his book *Introduction to the Logic of Psychological Measurement*: "According to the representational theory of measurement, the role of numbers in measurement is to represent. Numbers are assigned to empirical entities in such a way that certain relations between the numbers assigned represent empirical relations between the entities" (p. 29).

The isomorphism between these properties of the numeral series and certain empirical operations which we perform with objects permits the use of the series as a *model* to represent aspects of the empirical world (Stevens, 1946, p. 677).

The Stevens paper has exerted a major influence on the contemporary landscape of psychology and other social and behavioral sciences generally and has been incorporated into much contemporary statistical practice. Not only is his theory of scale types regarded as definitive by many within the behavioral sciences but it also stimulated interest in the study of scale types which eventuated in the development of the rigorous discipline of the axiomatic approach to *Relational Theory of Measurement* (RTM) from the 1960s until now. It could be said that Stevens' work with measurement scales represents an important contribution to the overall direction of the RTM, but not to the more thorough and conclusive *axiomatic* approach that is represented in the three volumes of *Foundational Measurement Theory* (Krantz et al., 1971; Suppes, Krantz, Luce, & Tversky, 1989; Luce, Krantz, Suppes, & Tversky, 1990).

1.2.2.4 Foundations of Measurement, the "Received View" of RTM

Narens (2002) gives a terse and cogent summary of the history of how the axiomatic approach to the relational theory of measurement and the three volumes of *Foundations of Measurement Theory* came to be, as illustrated in Figure 1.2.

> Measurement in its broadest sense consists of assigning numerically based entities to qualitative objects... This approach dates back at least to the foundational work on physical measurement by Helmholtz (1887) and Hölder (1901). It was also used by Hilbert (1899) to assign numerical coordinates to points of a qualitative Euclidean geometry. Much later, abstract versions of it for general measurement theory were given by Scott and Suppes (1958), Suppes and Zinnes (1963), and Pfanzagl (1959, 1968). The version of Scott and Suppes was later adopted as the theoretical basis for measurement in the seminal work on the subject, *Foundations of Measurement, Vols. I, II, and III* (Krantz, et al. 1971; Suppes, et al. 1989; and Luce, et al. 1990). Over the last three decades, this abstract version has dominated the theoretical measurement literature. For lack of better terms, I call this approach to measurement the *received representational theory*... (p. 211).

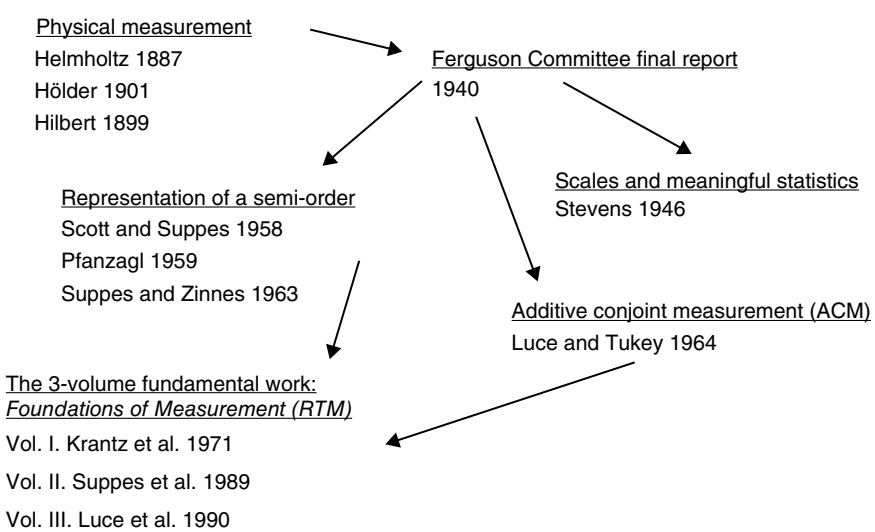


Figure 1.2 The development of RTM, the representational theory of measurement.

The early 19th century beginnings of measurement theory as the assigning of numerals to objects is clarified by Campbell's (1920, 1928) example of the isomorphism between empirical concatenation and numerical addition. Stevens takes a next step by defining the concept of measurement as "the assignment of numerals to objects *by rules*." He then makes an important strategic move by broadening the application of the concept with his statement that "the fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement." These three initial steps are followed by the axiomatic approach represented in the work of Scott and Suppes (1958), the work of Luce and Tukey (1964), and culminating in what Narens refers to as the dominant "received" view contained in the three volumes of *Foundations of Measurement* (Krantz et al., 1971; Suppes et al., 1989; Luce et al., 1990).

Quite abstractly, and with logical proofs and mathematics not accessible to many researchers in the behavioral or human sciences, this approach identified new classes of numerical structures that form the basis for measurement scales including more than the four types of scales popularized by Stevens. They then proved theorems of the existence of representations that mapped the numbers into one of the numerical structures thought to be appropriate to the empirical domain. They went farther and proved theorems of uniqueness that identified equally good parameters for representations into the mathematical structure, those that could handle data transformations as well. This is important. We don't want to hold our collective breaths to make sure we can use more than ordinal mathematics. We want at least additive structures so we can use least squares, correlations, factor analysis, and the like.

Patrick Suppes, the first author of the 1989 volume of Foundations of Measurement Theory, Volume II, earlier hoped this work would advance the field of practical calculations. Scott and Suppes (1958) remarked: "A primary aim of measurement is to provide a means of convenient computation. Practical control or prediction of empirical phenomena requires that unified, widely applicable methods of analyzing the important relationships between the phenomena be developed. Imbedding the discovered relations in various numerical relational systems is the most important such unifying method that has yet been found" (pp. 116–117).

Consider this spectrum that summarizes Scott and Suppes' two requirements for practical control, first, the use of widely applicable methods of analyzing the important relationships between the phenomena, and then that the discovered relationships be imbedded in various numerical relational systems. These are the two sides that need to be brought together, the empirical on the left and the numerical on the right. We have seen with Campbell's simple beam balance example how the empirical-to-numerical connection can be illustrated and made possible by concatenation:

Constructs and "would-be" scales from qualitative reality ↔ Imbedding (injecting)[9] the relationships found into a numerical system

A student of Suppes, Zoltan Domotor, wrote a chapter in the book *Philosophical and Foundational Issues in Measurement Theory* (Savage & Ehrlich, 1992). It cut deeply into the critique of axiomatic RTM and was given the position of the final chapter in that book. In it, he advances the position that of the two sides of the spectrum drawn above, the imbedding side cannot supply all of the needed theory. He called the left-side projection from theory. He argued that projection must take precedence and must precede imbedding, "...my conclusion will be that representation or injection of the qualitative into a quantitative is fully justified only when coupled with the dual

9 Scott and Suppes call this mathematical expression "imbedding" while Domotor calls it "injection."

procedure of projection" (p. 196, italics ours). By projection he means "The idea is that theoretical models are projected onto their observational structures..." (p. 195).

Domotor is framing an argument as a measurement philosopher. He does not give technical procedures for how to project theory onto an observational structure nor how to get either the theory or the models that tell us what to observe. Scott and Suppes in the quote above also did not give much attention to how we reach into the qualitative reality and extract theoretical constructs and turn them into scales. They talked about the use of "widely applicable methods of analyzing the important relationships" but then zeroed in on imbedding them into a numerical structure. Right from the beginning of his article, Domotor seems as committed to quantitative scales as Scott and Suppes but reminds us that both sides are needed.

Domotor gives his readers, including those who might be psychometricians, a strong statement of what needs to be done first. Psychometricians with broad experience developing instruments have learned how to identify and define empirical "constructs," find observable indicators of them, and thus create scales. As psychometricians, we could name the left side as "design and develop constructs and scales from qualitative reality." In Chapters 4 and 5, we will explain how analyzing empirical domains and developing domain theory can lead to models that can be used to "project," that is to design and develop instruments that tap into observational structures (Mislevy, Almond, & Lucas, 2004; Mislevy & Risconscente, 2005).

Psychological instruments are analogous to the extravagant instruments used at the Olympics but more modest in scope. It is useful to note that both types of instruments can provide the means for enabling users to "make contact" with tasks calibrated on scales where the instrument delivers a number that can be operated on by psychometric procedures. Also, the order of the tasks, or the order of physical or mental attainments in the theory, has a meaning that can be used in that all-important aspect of validation, "interpretation." That contact, together with enough such tasks for reliability, makes possible the issuing of a score after statistical work. The meaning comes from the theory and from the validating evidence that there is some reality in what people are doing with their minds or in their physical performances—an invisible mental or physical reality that can be captured by scores from an instrument.

Domotor (1992) objects to the empiricist approach and promotes a realist position. The theory is more than a language blessed with abstract terms to bandy about, nor is it just to make predictions. There is no reason to believe that the world contains anything that corresponds to the abstract terms of empiricism. Instead, he argues, "...a theory is more than just an instrument for driving predictions; it must satisfy our demand for an explanation by telling us how and why things happen" (p. 196). However, for this purpose, the theory must aim at a genuine description of what is happening. The practice of science shows us that it is precisely by viewing theories realistically that science progresses. It is by thinking of theories as literal albeit perhaps approximate descriptions of what goes on that we can suggest good ways of extending or generalizing them and can see where they might require modification.

The "received view" of RTM in the Krantz et al. three volumes has not been without comments, suggestions, and critique, as exemplified in Domotor's call for a realist approach to understanding measurement. Others have pointed out and objected to the proliferation of axioms, proofs, and theorems in the Krantz et al. version of RTM that in and of themselves lack practical utility. Domotor's pithy comment on this, after giving 12 pages explaining logical reasons for the "troubling profusion of representation results" was, "I suggest to project first and represent later" (p. 214).[10]

10 This is an intriguing thought that perhaps more clarification on the "qualitative reality" side could simplify the task of proving numerical representation and uniqueness.

However, what such theorems do is provide foundational assurance that our number system assumptions are justified. It would be foolish to not recognize the achievement of the pioneers who have given the world Foundations of Measurement Theory, but this work was not designed to replace the tried-and-true methods of true score test theory. That possibility has been suggested by some (Michell, 1990, 2020), but as Cliff (1992) observed 30 years ago the expected axiomatic RTM revolution in psychometric theory and methods had not occurred by 1992, and 30 years later it still hasn't. The discipline of psychometrics continues to move forward with practical and useful test theory methods for the most part unaffected by RTM. We should be patient, however, with those who continue their difficult RTM work in hopes of possible useful convergences.

Fortunately, the axiomatic RTM theorists are self-critical. Luce and Narens (1994) published the results of research funded by a National Science Foundation Grant. In it, they listed ongoing work, various unsolved problems with RTM, and several suggested new directions. They listed 15 problems requiring additional work. Among them are two of the most important issues for our key interests in this book: the practical work of developing actual measurement scales and clarifying the meaningfulness issue.

One useful and practical result to come from mathematical RTM is Luce and Tukey's (1964) demonstration of the power of ACM. When its axioms are met, it produces an interval scale without the operation of concatenation. If an instrument can be shown to meet the axioms of ACM, and the mathematical model fits the data well enough to assume conditional independence, investigators have evidence that they are justified to treat their constructed scale as additive with a convenient zero and a unit that is as persuasive as the unit and zero created for the Fahrenheit and Celsius temperature scales.

1.2.2.5 The Contrast Between Foundations of Measurement and the Science of Applied Psychometrics

At the end of Section §5, we indicated that we would only touch lightly on the "logical, theoretical, and philosophical foundations of measurement." We took a brief look at the history and principles of representation and meaningfulness as embodied in the RTM and more specifically, the axiomatic approach to the RTM as it emerged in the mathematical psychology literature of the past 50 or 60 years.

This is not to be confused with CTT and its derivatives, which began much earlier—at the beginning of the 20th century—which constitute the foundational practical science upon which the past century of psychological testing is based. With our brief look at RTM now completed, we turn to this dominant approach to psychometrics throughout the remainder of the book—so-called "true score theory," which includes CTT, item-response theory, and the myriad methods that have grown out of them. The axiomatic approach to RMT has had very little impact on psychological testing or psychometric practice. It consists primarily of logical clarification, conceptual demonstrations, and axiomatic logical proofs which arose in mid-century as an effective rebuttal to the accusation from some quarters of the scientific community that measurement in psychology, and indeed in all the behavioral sciences, is at best unscientific and at worst perhaps meaningless.

Each of these two traditions is referred to as "measurement theory" as shown in Figure 1.3 but they differ substantially in their approach, in their historical roots, and in their practical impact. As shown in that figure, both kinds of measurement theory have both a theory aspect and also a scaling aspect. The measurement theory aspect is foundational and formal and deals with the mathematical and conceptual/theoretical conditions under which scales can be constructed, while the actual process, the practice by which numbers are assigned to objects, or attributes of objects, is called scaling.

	Numerical and philosophical foundations	Focus on application and scale development
1. True score theory - Spearman (1904a, 1904b, 1907, 1910, 1913) - Thurstone (1931) - Gullikson (1950) - Lord and Novick (1968)	**Quadrant A** Latent trait theory Classical test theory (CTT) Item response theory (IRT) Theory of mental testing	**Quadrant B** Reliability measures Validity measures Heterogeneity measures Item analysis methods
2. Foundations of measurement - Scott and Suppes (1970) - Luce and Tukey (1962) - Krantz et al. (1971, 1989, 1990)	**Quadrant C** Representation theorems Uniqueness theorems Meaningfulness theorems	**Quadrant D** Additive conjoint measurement Rasch scaling as ACM Prospect theory

Figure 1.3 A comparison of two contrasting approaches to psychological measurement: (1) True Score Theory, the traditional psychometric testing tradition, with its myriad derivative developments; and (2) the Foundations of Measurement axiomatic approach to the RTM, the current "received view" of RTM.

True score theory, Quadrant A in the table, is essentially the main line of theory and practice in psychometrics as it has grown over the past 12 decades. It includes the majority of what is explained in the remainder of this book. It begins with the emergence of factor analysis at the beginning of the 20th century and the subsequent developments of the theory and mathematics of measures of reliability, validity, test homogeneity, and a wide variety of other psychometric equations and processes—fueled by Pearson's (1896) correlation coefficient, his discovery of the mathematical basis of factoring (2001), and Spearman's introduction of factor analysis. Most of these date back to the first half of the 20th century. These historical developments are described well by Gulliksen (1950), with many references to early fundamental sources such as Spearman (1904a, 1904b, 1907, 1910, 1913), Thurstone (1931), and many others. Gulliksen acknowledges that his book draws heavily upon the teachings of Thurstone "and particularly upon his now unavailable book *Reliability and Validity of Tests* (Thurstone, 1931)."

Latent trait theory is also a part of Quadrant A and includes the early sources on factor analysis such as those by Thurstone (1935, 1947), Harman (1960), and Horst (1965), that chronicle the development and elaboration of a wide variety of factoring models. These early psychometric efforts are referred to collectively as CTT (classical test theory) to distinguish them from modern test theory, that is, item response theory (IRT). CTT as described in exemplary texts such as Lord, Novick, and Birnbaum (1968) and Allen and Wen (1979) was first codified by Novick (1966). The definitive source explicating and summarizing these quantitative foundations of psychometrics over the first 70 years of its existence is the 1968 Lord and Novick text *Statistical Theories of Mental Test Scores, with Contributions by A. Birnbaum*. Probably no other single book since that time has had a comparable impact upon the past century of psychometric practice. It is a one-volume sourcebook on the statistical/mathematical basis of CTT supplemented by an introduction to the beginnings of item-response theory in Chapters 17–20, written by Allan Birnbaum.

Quadrant B consists of the many psychometrically useful analytical practices and tools that grew out of the numerical and philosophical foundations of CTT and IRT, such as those presented in the Gulliksen book or the Lord and Novick book, including Jöreskog's (1969) epoch-making maximum likelihood method of confirmatory factor analysis (to be covered in Chapter 6 and Computational

Guide 1). It also consists of the multitude of specific psychological tests and measures that have been produced using these practices and tools. This would include the over 14,000 specific testing instruments chronicled in the 21 published editions of the *Buros Mental Measurement Yearbooks* from their inception in 1938 down to the most recent edition in 2021 (Carlson, Geisinger, & Jonson, 2021). This was the golden age of test development, still in progress.

Quadrant C is the axiomatic approach to the RTM. It began in the 19th and early 20th century, as was presented in Figure 1.2 and the accompanying pages, and culminated in the three volumes of *Foundations of Measurement,* the currently accepted method of RTM. This approach had its beginnings in the mathematical psychology literature of the 1960s and 1970s (Coombs, Dawes, & Tversky, 1970), as a natural follow-up to the Ferguson committee controversy, Stevens' gentle rebuttal in his 1946 measurement scales paper in *Science*, and the surprising and convincing demonstration by Luce and Tukey (1964) of rigorous psychological measurement without concatenation using simultaneous conjoint measurement. Stevens' four measurement scales have become second nature to the behavioral science disciplines and could be considered to also be a variety of RMT (the Relational Theory of Measurement), but not of the *axiomatic* approach to the Relational Theory of measurement that now characterizes the mathematical psychology literature. Clearly, Stevens' four measurement scales do not constitute a rationally defensible justification of the fundamental principles of measurement but they have been a useful part of psychometric progress.

In addition to ACM, the poster child of axiomatic RTM, there have been many Quadrant D mathematical models loosely connected to axiomatic RTM developed over the past century, including signal detection theory (SDT), other psychophysical models, information theory, unfolding theory, mathematical learning theory, scalogram analysis, multidimensional scaling, Thurstone's law of comparative judgement and his law of categorical judgement, models of preferential choice, utility theory, etc. (Coombs, 1964; Torgerson, 1966). Perhaps one of the most notable achievements to grow out of axiomatic RTM was the behavioral economics work of Tversky and Kahneman. Daniel Kahneman was awarded the Nobel Prize in Economics in 2002 for his part in the development of prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1974), just six years after the death of his research partner Amos Tversky, in 1996. A brief axiomatic treatment of prospect theory was given in the appendix of Kahneman and Tversky's (1979) paper, with acknowledged assistance from David Krantz, the lead creator of the three foundations of measurement volumes. Perhaps there is also a tie-in between axiomatic RTM and Rasch (1960) scaling, a precursor of item response theory. There is some evidence that the Rasch scale fulfills the conditions of ACM (Perline, Wright, & Wainer, 1979; Karabatsos, 2001; Newby, Conner, Grant, & Bunderson, 2010).

1.2.2.6 Metrology and Measurement Units

While the Ferguson Committee was engaged in its seven-year debate, failing to consider forms of measurement other than those used fruitfully by physicists, engineers, and chemists, International Metrology[11] was becoming more and more international, and developing cooperation on setting standards and promulgating a System of Units (SI), a metric system that scientists of all stripes could use worldwide. It led to great economic and social gains, through the application of science-based standards throughout industry and the professions. Metrology is the science of

11 International metrology has been with us for over 200 years. The early metrologists received a major boost for their "meridian-arc" definition of the meter as one ten-millionth of the earth's circumference following the completion of Méchain and Delambre's six-year survey of the distance from the equator to the North Pole in 1798 (Ferreiro, 2011; Alder, 2021)

measurement. As philosophy could be considered "thinking about thinking," metrology could be considered "measuring measurement" (Brown, 2021). It is much more mathematical than other approaches to measurement. It has three divisions, Scientific Metrology, Applied (or industrial) Metrology, and Legal Metrology. It has made remarkable advances up through 2019, during which time all physical artifacts for measuring length, and finally in 2019, calibrating weights, were replaced. Physical artifacts are not precise enough—not easily calibrated nor traced. International Metrology is not usually explained in textbooks of this sort, but it is so important to the future of any scientific field that we have prepared a Chapter 8 entitled "Metrology: Scientific, Applied, and Legal.

In this book's introduction, we promised to show that while a demonstration of concatenation is *sufficient* to establish a measurement scale, it can be shown to not be a *necessary* condition for scientific measurement. This is most clearly shown in the stunning achievements of contemporary metrology scientists. In May 2019, the International System of Units (SI) replaced the last physical object suitable for concatenation demonstrations. It was THE KILOGRAM (called the IPK [International Prototype Kilogram]) that was carefully stored under ultra-safe and controlled conditions in a metrology laboratory in Paris. It has been retired and replaced by an average of new international estimates of fundamental constants, following the latest experimental estimates of the Planck constant, averaged from nine Metrology Laboratories in several countries to a relative uncertainty of 1.2×10^8 decimal places.

This is a profound achievement in the broad science of measurement, retaining the familiar physical measures such as meters and grams, but tying them to seven fundamental constants of the physical universe. And it comes with a refreshing openness and encouragement for new approaches to measurement in every conceivable domain. Who can imagine what its effects might be on the fledgeling science of psychometrics in the coming decades?

You can read more about this amazing attainment in worldwide cooperation in Stock, M., Davis R., de Mirandes, E., and Milton, M. J. T. (2019), "The revision of the SI—the result of three decades of progress in metrology." The abstract of this *Metrologia* paper reads as follows:

> On 16 November 2018, a revision of the SI was agreed upon by the General Conference on Weights and Measures. The definitions of the base units were presented in a new format that highlighted the link between each unit and a defined value of an associated constant. The physical concepts underlying the definitions of the kilogram, the ampere, the kelvin and the mole have been changed. The new definition of the kilogram is of particular importance because it eliminated the last definition referring to an artefact. In this way, the new definitions use the rules of nature to create the rules of measurement and tie measurements at the atomic and quantum scales to those at the macroscopic level. The new definitions do not prescribe particular realization methods and hence will allow the development of new and more accurate measurement techniques.

Figures 1.4 and 1.5 are photocopies of tables 1 and 2 taken from page 128 of BIPM (2019), the Bureau International des Poids et Mesures (English text version in Newell & Tiesinga, 2019, *The International System of Units (SI), ninth edition*). Figure 1.4 shows the seven standard constants, together with the symbols, numerical values of the constants, and the defining unit of each. Figure 1.5 gives the base quantity of each of the seven familiar base quantities derived from each constant, the typical symbol, and the base unit name of the measuring units for each (second, meter, kilogram, etc.) with its symbol.

Defining constant	Symbol	Numerical value	Unit
Hyperfine transition frequency of Cs	$\Delta\nu_{Cs}$	9 192 631 770	Hz
Speed of light in vacuum	c	299 792 458	m s^{-1}
Planck constant	h	$6.626\ 070\ 15 \times 10^{-34}$	J s
Elementary charge	e	$1.602\ 176\ 634 \times 10^{-19}$	C
Boltzmann constant	k	$1.380\ 649 \times 10^{-23}$	J K^{-1}
Avogadro constant	N_A	$6.022\ 140\ 76 \times 10^{23}$	mol^{-1}
Luminous efficacy	K_{cd}	683	lm W^{-1}

Figure 1.4 The seven defining constants of the *SI* and the seven corresponding units they define, a photocopy of table 1 from page 128 of *The International System of Units (SI)*. This SI Brochure is distributed under the terms of the Creative Commons Attribution 3.0 IGO (https://creativecommons.org/licenses/by/3.0/igo/).

Base quantity		Base unit	
Name	Typical symbol	Name	Symbol
Time	t	Second	s
Length	l, x, r, etc.	Metre	m
Mass	m	Kilogram	kg
Electric current	I, i	Ampere	A
Thermodynamic temperature	T	Kelvin	K
Amount of substance	n	Mole	mol
Luminous intensity	I_v	Candela	cd

Figure 1.5 The seven *SI* base units and their names and symbols, a photocopy of table 2 from page 128 of *The International System of Units (SI)*. This SI Brochure is distributed under the terms of the Creative Commons Attribution 3.0 IGO (https://creativecommons.org/licenses/by/3.0/igo).

1.2.2.7 Measurability in the Behavioral and Social Sciences

Psychologists often speak of nomological nets that provide linkages among a variety of theoretical constructs. This attainment in the realm of physics and the multiplicity of types of experiments in the various metrology labs give a deep example of how advanced nomological nets can become, and how long it can take, even in physics. Concatenation can still be performed because each country can create its own standard prototype kilogram—or any other unit. But it cannot be said that a demonstration of concatenation is necessary to prove the measurability of any standard unit in the SI. The references above also provide a very clear common-sense understanding that physical phenomena that have no volition and answer only to unchanging natural laws can be measured with such incredible precision that it would be ludicrous to require that standard of accuracy to phenomena in the human-created domains wherein behavioral measurement takes place.

Often only order is necessary to have explainable theories that are consistent and useful, and are backed by solid constructs, good instruments, and good implementation techniques to reduce construct-extraneous sources of error. But as shown in the axiomatic RTM sub-section earlier, with ACM, when people's scores and a set of tasks are conjoined, we can establish additivity as well. But, ratio scales are not always to be expected nor needed in human-made domains that possess the numerical values appropriate for instrument development.

1.2.3 (§7) *Scientific* and *Stochastic* Variables: When May "Data" Be Treated By Statistical and Test Theory Methods?

The genius of CTT is to conceptualize test scores as consisting of two components: the "true score" which can be thought of as the deterministic component, and "measurement error," which can be thought of as the stochastic/probabilistic component—a random variable. Random does not mean haphazard—far from it. A random sample, for example, is rigorously representative of the population from which it is drawn. That is, each member of the population has an equal chance of being included in the sample. It is an exacting process to create a "sampling frame," a list that includes all in the population, and to then use a truly random process, something like a random numbers table, to select the sample.

Stochastic variables can provide an important kind of leverage. A random variable, although unpredictable in a single observation, is predictable in the aggregate. Suppose, for example, you wish to stochastically model the distribution of the outcomes, the sums of values, in tossing two dice. In other words, you wish to create a *probability model* of outcome sums. Obviously, one could not predict any particular outcome of tossing two dice, but if the dice are truly unbiased and random, one could specify with precision and exactness what the distribution of outcomes would be across many such tosses, as shown in Figure 1.6 which constitutes a kind of "diagrammatical demonstration." Each of the 36 possible rolls of the 6 possible outcomes on one die and the six on the other have an equal probability of occurring. Each of these possible, and equally likely, two-dice rolls for each sum value are grouped above the appropriate sum value, creating a systematic distribution. Each of these 11 sum values (from "2" to "12" inclusive) has a precise probability as shown in the probability vector above the stacks of outcomes. These could be considered to be population outcomes, but they are more than that. They are ontological probabilities that capture the true and timeless nature of the mathematics of dice tossing for samples of size two.

What has been created here is a *sampling distribution of sums*. It is essentially equivalent to a sampling distribution of means (SDM), except that we have not divided the sums of the two-dice tosses by N. If we now had the time, the inclination, and the patience, we could create similar sampling distributions for three-dice tosses ($N = 3$), which would have sixteen unique sum values along the x-axis, the abscissa (with sum values ranging from 3 to 18), and 216 three-dice outcome permutations; or even for four-dice tosses ($N - 4$), which would have an abscissa with 21 unique sum values (4–24), and 1296 four-dice outcome permutations. And, even though this sampling distribution with $N = 2$ is triangular in shape, with increasing sample size (N) from 3 and up, the sampling distribution of sums will become more and more normally distributed, that is, more Gaussian. This is in accordance with the *Central Limit Theorem,* which holds that (regardless of the shape of the

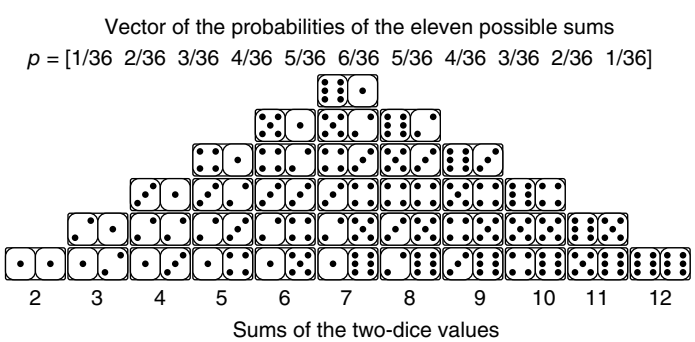


Figure 1.6 A graph of the sampling distribution of the 36 possible paired-dice tosses and their respective probabilities.

population distribution) as sample size increases the sampling distribution of means, the SDM, will converge to Gaussian. The same is true of the sampling distributions of sums and of proportions.[12] In other words, a bell-shaped Gaussian distribution is the limiting form. In Appendix 2, methods are introduced for calculating descriptive statistics for probability models. Means, variance, and other statistics can be used to describe with statistical precision the nature of distributions, spatial patterns of probability densities, and so forth. CTT consists of a variety of ways of managing and reducing error and discovering "true score" structure when an error is present. When applied to measurement error, rather than dice tossing, probability models enable us to characterize, estimate and reduce measurement uncertainty with precision as demonstrated in Section §13. Hence the value of stochastic modeling in psychometrics.

A theoretical probability model, like the one just described, has three necessary conditions: first, the outcomes in the model are *mutually exclusive*, second, the outcomes are *exhaustive*, and third, the *probabilities sum to 1.00*. In Figure 1.6, the outcomes are clearly mutually exclusive: if the sum of two dice is 7, it cannot also be 8 or 9. The outcome possibilities are also exhaustive: there are no other possible sums of two dice than the eleven sums specified. The reader can also verify that the eleven probabilities in the probability vector above the diagram do indeed sum to one. This sum of 1.00 is another way of stating the property of *necessity*—each one of the 11 sum-value outcomes and their associated probabilities are necessary to the complete model. Given the verification of these three properties, we can conclude that this sampling distribution of sums is indeed a probability model.

The important part that random variables play in psychometric theory and CTT will not be fully explained here but will become clearer as we progress throughout the remaining chapters. This is enough to establish that there is indeed explanatory power in establishing a variable as a stochastic variable, a true random variable.

1.3 Statistical Methods: Location, Dispersion, and Statistical Inference

1.3.1 (§8) Group Level of an Attribute—The Mean Location of the Group's Scores on an Interval Scale

We will now demonstrate the calculation of indices of central location and dispersion of scores using fictional quasi-Galtonian data as if measured in a sample of N pairs of fathers and their firstborn sons. If we accept the procedure of measuring head circumference to be a scientific variable recognized as a stochastic variable x, then its representation conditions enable measurement on an interval-metric scale. Hopefully, it can be seen then that the arithmetic *mean* is a very informative characteristic, at least intuitively, if one wants to compare the overall levels of two groups of people in two data sets of a purely quantitative nature. Besides the arithmetic mean, there are other useful characteristics that express location—the median and the mode. All three are expressive of the overall tendency of a data set to be "located," or its position to be "centered," somewhere on a scale. So, quite naturally they are referred to in statistical terminology as *characteristics of location* or *measures of central tendency*.

Consider a data set of fictional head circumference data for a sample of $N = 30$ fathers and also for their sons (one son each, so also $N = 30$). Suppose that we also round the continuous-value head

12 This is not only true of means and sums, but also of proportions, since the proportion is a special case of the arithmetic mean. A proportion is the mean of a binary variable. Note, however, that the rolling of dice or the tossing of coins creates discrete distributions, while the Gaussian distribution by definition is continuous. However, although the sampling distribution created by coin tosses is technically a discrete binomial distribution, it is closely approximated by a Gaussian distribution as N becomes large. (See Brown, Hendrix, Hedges, and Smith (2021), pages 43 to 58, the section entitled "Classical probability theory and the binomial: The basis for statistical inference.")

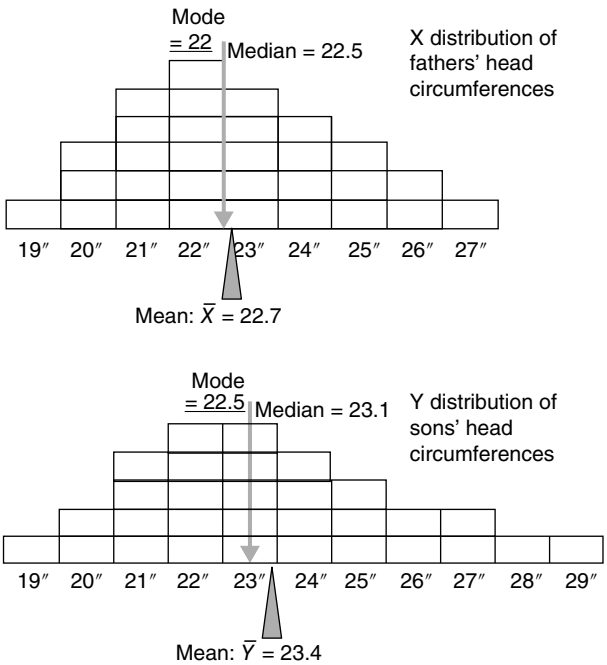


Figure 1.7 Histogram comparison of fictional head circumferences of 30 fathers with corresponding fictional head circumferences of their sons, with the mean, median, and mode shown for each distribution.

circumferences to the nearest inch. The values measured are pictured in inches as points on an interval scale in Figure 1.7. The example will be simplified and more illustrative working with rounded discrete data. If we represent the data for fathers as 30 wooden blocks, each block representing the head circumference value of one of the 30 fathers, and each stacked above its appropriate integer-rounded circumference value along a numerical scale from 19″ to 27″, as shown in the upper half of Figure 1.7, it creates a *histogram* revealing the shape of the distribution of head circumferences for fathers. The same is done for the head circumferences of sons, as shown in the lower half of Figure 1.7 Each value within each distribution has a "relative frequency," or probability. For example, the value 27″ in the fathers' distribution contains only 1 of the 30 blocks, so its relative frequency or "probability"[13] is 1/30. Similarly, the value 24″ has a probability of 4/30 within this sample, and the value 22″ therefore has a probability of 6/30, etc.

We can find the average value (also called the *mean*) of the 30 scores of the distribution for the fathers, by first summing all thirty of the observed values, which comes to 681, and then dividing the 681 by 30, to get a mean of 22.7.

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$\overline{X} = \frac{\begin{(19 + 20 + 20 + 20 + 21 + 21 + 21 + 21 + 21 + 22 + 22 + 22 + 22 + 22 + 22 \\ +23 + 23 + 23 + 23 + 23 + 24 + 24 + 24 + 24 + 25 + 25 + 25 + 26 + 26 + 27)}}{30}$$

$$= \frac{681}{30} = 22.7$$

13 We are using the term *probability* descriptively here, rather than ontologically, to simplify the demonstration of the conceptual basis of the calculation of fundamental statistics.

We use the symbol "X-bar" $(\overline{X})$ to represent the mean of distribution X. The capital sigma sign (Σ) in the numerator of the formula has an "$i = 1$" at the bottom of the sigma and an "N" at the top of the sigma. These indicate that you are to sum all the X_i values, from the first (with subscript = 1) to the last (with subscript = N). This sum is then divided by N. A simpler notation, and a more efficient way to get the mean for this sample is to obtain the "sum of products" of each of the X_i values multiplied by its own probability p_i, as explained in Appendix 2.

$$\begin{aligned}\overline{X} &= \sum(X_i p_i) \\ &= 19\left(\frac{1}{30}\right) + 20\left(\frac{3}{30}\right) + 21\left(\frac{5}{30}\right) + 22\left(\frac{6}{30}\right) + 23\left(\frac{5}{30}\right) + 24\left(\frac{4}{30}\right) \\ &\quad + 25\left(\frac{3}{30}\right) + 26\left(\frac{2}{30}\right) + 27\left(\frac{1}{30}\right) \\ &= \frac{19}{30} + \frac{60}{30} + \frac{105}{30} + \frac{132}{30} + \frac{115}{30} + \frac{96}{30} + \frac{75}{30} + \frac{52}{30} + \frac{27}{30} = \frac{681}{30} = 22.7\end{aligned}$$

The mean of Y, the distribution of the head circumferences of the sons, can be similarly obtained.

$$\begin{aligned}\overline{Y} &= \sum(Y_i p_i) \\ &= \frac{19}{30} + \frac{40}{30} + \frac{84}{30} + \frac{110}{30} + \frac{115}{30} + \frac{96}{30} + \frac{75}{30} + \frac{52}{30} + \frac{54}{30} + \frac{28}{30} + \frac{29}{30} \\ &= \frac{702}{30} = 23.4\end{aligned}$$

Calculations such as this can also be done somewhat more elegantly and systematically using matrix multiplication, as shown in the first two paragraphs of Appendix 3.

The mean is an agglomerative characteristic of data distribution, with two interpretations. The first is empirical, as a joint level of an attribute (such as head circumference) that is common to a group of people. That is, it summarizes the central tendency location of the group. The second interpretation is formal, as a point representing a mathematically central location of a set of numbers on a scale. The mean can also be thought of as the "expected value," in the sense that it gives the value that is "expected," that is, most likely for an unknown score, as explained in Appendices 1 and 2. It can also be thought of as the balance point for the distribution of scores as can be observed for the two distributions shown in Figure 1.7, where a small "wedge" or fulcrum is shown at the location of the mean for each of the distributions. If we were to model one of these distributions as an actual physical object with 30 wooden blocks stacked on a ruler or scale of negligible weight, the mean would be the precise location where a wedge must be placed to perfectly balance the distribution. As such, the mean is obviously highly affected by extreme values. If, for example, you were to take the block at location "27" on the X distribution, the highest score, and move it out to a location of "36," it would shift the balance point, the mean, from 22.7 to 23.0.

A typical textbook list of major descriptive characteristics of the central tendency location of a data set includes, besides the mean:

the *median*, defined as the middle value of the scores when ordered by magnitude, that is, as the point along the scale that bisects the area of the curve; and
the *mode*, the value with the highest frequency, so, in a sense, its "typical" value.

The mode as a physical characteristic of the stack of blocks would be the location of the highest stack of blocks, which for the X distribution is 22, with a frequency of 6. For the Y distribution, there

are two scores, 22 and 23, that each has a frequency of 5, so the mode is in the center of them, with a value of 22.5. See Appendix 1 for an account of the properties of mean, median, and mode, and their calculation.

1.3.2 (§9) Comparing Median vs. Mean: Which One Is "Fair"?

Let's go back and compare the mean $\overline{X}$ vs. the median with respect to their differing behavior when the data undergo changes or move along the scale. The mean, as the sum of the individual values divided by N, takes the magnitude of every score into account, whereas the median does not. The mean reacts to any small change even in one value—as we can recognize by recalling the formula again:

$$\overline{X} = \frac{(x1 + x2 + \cdots + xi + \cdots + xN)}{N}$$

Therefore, the mean as a measure of central tendency operates as a center of gravity of the data points and their distribution along the scale. So, the mean could be "computed" mechanically by balancing a measuring stick to equilibrium when the negative deviation of the data points to the left are counterbalanced by positive deviations to the right. Here we have an analogy with the first mechanical moment—balancing the data set on the scale to equilibrium means that the scale, as a "lever," is supported at the data center so that the sum of the deviations in magnitude from left to right, the first-order central moment, is zero.

Perhaps the greatest advantage of means as a measure of central tendency is that they fit a Gaussian distribution as the sample size N over which they are summated becomes large. This is the *Central Limit Theorem* discussed in Section §7. The *Central Limit Theorem* does not apply to medians.

The median is not as sensitive and does not respond to small moves of an individual data point until it jumps over the original median to the other side, i.e. only if the value changes its ordering relation with respect to the original median. Therefore, we could say that the median is not so "fair" as to represent the influence of each individual data value in relation to the overall central characteristic of the data set: Imagine how a change in one subject's individual attitude can influence the overall attitude of a group of people if the "composite group attitude" were measured by the arithmetic mean on the one hand or by the median on the other. The insensitivity of the median to individual data points is sometimes useful as it is not so influenced by just one or a few extreme values falling far outside the typical range. Measuring central tendency with the median rather than the mean could be a reasonable policy in some particular areas of research. For example, in reaction time studies there are occasionally extreme outliers when a distraction can make one reaction time many times larger than the rest. This could create misleading comparisons if means were used. The problem could be avoided by using medians. In statistical terminology, it is said that the median is robust with respect to outliers. It can also be useful to compare the mean and median of the same data set to learn what happens when a certain non-typical trial or nontypical subject (or a couple of subjects or trials) is removed.

1.3.3 (§10) A Psychometric Look at the Mean $\overline{X}$, Item "Difficulty," and p

Many behavioral and psychological attributes of people as subjects in a research study are observed through diagnostic methods that are psychological "tests" in the narrower meaning of the usual common-sense usage, such as admission tests, etc.—*not* in the broader sense, which calls *any*

diagnostic method a test in the special language of Test Theory. A lot of the "real" tests may be called "performance tests" in a general sense as we speak about academic performance in our rather competitive civilization. Such a test, whose scores are formalized as research variables and properly represented by corresponding random variables X, or Y, ..., typically stands for a real practical task to be solved by the respondent, and the *test score*, x or y, then expresses a quantitative level of "how well the test task has been performed." That is, it is an index of the subject's performance. Then, unlike the mean of somatic measurements in Galton's studies, a mean of the scores of such a test y may also be interpreted as an overall *level of performance* $\overline{Y}$ of a particular group of persons. Also, we can compare the means of two groups of people with this underlying performance interpretation. For example, we can compare "performance" in a simple clinical test of verbal ability, let's call it the *V-test*, in two groups of patients with two different medications for dementia. Let's denote the score on the *V-test* by the random variable Y—it consists of the task to produce as many words as the patient is able in one minute, namely the words which start with a given letter, or names of animals, and so on.

In the history of psychometrics, the oversimplified explanation of a *performance score* was based on two concepts: the *ability* of the tested subject and the *difficulty* of the task he or she was supposed to perform. If the task has only two alternatives, performed or not performed—it means accomplished or failed—the "equation" of such a performance contained just a single *dominance relation* between the two members:

$$\begin{aligned} &\text{Abilitylevel} > \text{task difficulty} \quad \Rightarrow \quad \text{Accomplished, that is, performance is "good"} \\ &\text{Abilitylevel} < \text{task difficulty} \quad \Rightarrow \quad \text{Failed, that is, performance is "bad."} \end{aligned}$$

But even quantitative scores of the individual's performance are related to the difficulty level of the task. Performance and difficulty are complementary and counterbalance each other. Therefore, finding that Group 1 has a higher mean score than Group 2, that is, $\overline{Y}_1 > \overline{Y}_2$, could be interpreted as the higher ability level of Group 1, but also as an indication of various conditions which might have decreased the level of test difficulty in the testing session for Group 1.

Of course, the rigorous conditions of standardization should not allow substantial influence on test difficulty, but the test could be more or less difficult for given categories of population specified by age, sex, etc. So, it is important to remember that an unspecified "mean level $\overline{Y}$" can be interpreted in two ways: namely either as an agglomerate *attribute of the group of people* as related to a given test or as an attribute or *property of the method—of the test* with respect to the given sample or population.

The usual clinical viewpoint of "mean level $\overline{Y}$" evaluates and compares particular respondents through their test score representations considering them as attributes of *human qualities* that need to be treated, and possibly improved. On the contrary, from the psychometric point of view, the same raw materials, namely the test scores, are used to study and compare various modifications of a test and its properties as *characteristics of a diagnostic procedure* that are to be improved to offer better practical and scientific information.

Thus, by having a standardized test administered on a representative and large sample or, still better, after replications on several such samples, we can learn about special psychometric indices as indicators of the overall diagnostic quality of the test. Such a *test calibrating study* can yield important information about the psychological test or test battery and suggest practical modifications for its diagnostic improvements.

Information about the *appropriate difficulty* of a test for a certain category of the population, i.e. age, gender, education, etc. is one of the basic properties of any test. In its simplest case, the mean

score, $\overline{Y}$ is used as a *psychometric index* of *test difficulty*. It is one of the important diagnostic properties of the test.

Later we will introduce several other indices of additional psychometric properties. These will be more complicated and more focused on their use in psychometric analysis. But even in this very simple case, we can see the important relationship between test difficulty and *test validity*. Not yet defining what "validity" is or should be, let's rely on common sense, saying that validity is the theoretical, as well as the practical usefulness of a test, to diagnose or measure the desired behavioral attribute. Perhaps it is intuitively clear that if test difficulty is either too high or too low, its score is less helpful for distinguishing between people within a given population. Either all measured subjects would be at the "floor" score level, or respectively, close to the "ceiling" score level.

Now, assume for a while that the practical purpose of a given test is just to distinguish between acceptable and non-acceptable subjects. As common sense would suggest, the best and most appropriate difficulty level would be if the test's cutting point were at "fifty/fifty" or the 50% level, having the scale cutting point of the test calibrated for the given population. If the test score is quantitative, such a test should be calibrated so that the overall conditions determining the task difficulty are adapted so that the cutting point of the score is located at the mean of the population. In the example of the *V-test*, we can find such a modification by instructing respondents from a population of very old and undereducated persons to produce words starting with the letter *P*. This would certainly be easier for this most frequent first letter in English than to produce words starting with *X* or *Q*, for example. Also, we could use a time interval longer than one minute to make a test more accessible. Such manipulation can yield a version of the original test that will have its mean satisfactorily close to the cutting point that had been established by experts, say for diagnosing a "normal" level of performance.

If a test takes on the "accomplished/failed" or "good/wrong" or similar dichotomized scores, then its psychometric index of difficulty is directly given by the corresponding percentage of persons who passed or failed. Such tests usually are only parts of a battery or inventory of tests, and traditional psychometric terminology calls them *test items* or simply *items*. Scores of such items take on only two possible values, therefore those are often called *binary* or *dichotomous items*, with values $y = 1$ and/or $y = 0$ for coding scores, usually by 1 for "right" and 0 for "wrong." Although this kind of assigning numbers seems to be an arbitrary symbolization, it still has an important property. It could be called *quasi-quantification*. Such an artificial random variable created by this (0, 1) coding of values is also well known as a *dummy* variable. It has certain quasi-quantified properties since in certain mathematical models it allows us to treat such tests and items *as if* they were measurable interval-metric quantities.

One of those mathematical methods is the *mean of a binary item*: $\overline{Y} = p$. The mean is now identical to the *proportion p* of unities "1" in the set of *N* subjects who successfully completed the test task. Since $\sum Y_i$ is the number of "1" values, when it is divided by the sample size *N*, it produces a proportion, $p = \dfrac{\sum Y_i}{N}$. Correspondingly, $q = 1 - p$ is the proportion of those who failed, and it follows that $p + q = 1$. Of course, the higher proportion of those who passed successfully, i.e. the higher the *p*, the easier the item is for a given population. But, strangely, perhaps due to a century-long tradition, so-called "item analysis" still calls *p* "the *difficulty* of a test item." We must just passively accept this illogical name, which by common sense should rather be called item "easiness," with an excuse that it is simply a technical term and a definition. Some people call *p* the "index of difficulty," perhaps an index with an opposite direction of meaning, but perhaps we could avoid the conflict by using intuition.

1.3.4 (§11) Intra-group and Inter-individual Variability—*Dispersion* of Scores

A more detailed look at Figure 1.7 suggests that the central tendency information does not tell the whole story of how the head circumferences of fathers compare with those of their sons. The overall level of the attribute and location of the group on the scale is not enough. Not only is the midpoint of the head circumference distribution for sons slightly higher than the midpoint for the fathers' distribution, but the sons' data distribution also has a somewhat larger spread, and a slightly larger dispersion.

Perhaps the two groups, the fathers and the sons, could be almost of the same mean level but still differ in their spread, that is, their variability. The first intuitive look, then, could be at the maximum and minimum values in each group. This is also visually apparent in Figure 1.7.

Returning to the two histograms we can use the formula and calculate the range for the group of fathers and the range for the group of sons in that figure, to verify numerically that the sons do have greater variability in head circumferences. The range for the fathers is:

$$Ra_f = max_f - min_f = 27'' - 19'' = 8''$$

And the range for sons is:

$$Ra_s = max_s - min_s = 29'' - 19'' = 10''$$

This contrast is also visually apparent in Figure 1.7.

The range $Ra(x)$ is a characteristic of variability, while the median characterizes location. Although each represents a different aspect of the distributions, they both still have a certain mathematical property in common. Like the median, the range is insensitive to small changes in individual values, until they jump over its border. However, the range can be overly sensitive to an extreme value of an atypical individual in the group whose value is outlying far to the side of the group. Such an *outlier* causes the information about variability to be distorted. Therefore, a better and "fairer" characteristic would be based on similar mathematical properties to those of the arithmetic mean so that even a very small change of any individual value influences the overall characteristic.

The idea that the larger the overall deviations of values from their center, the larger the dispersion, suggests that the arithmetic *mean of the deviations* can be such a "fair" characteristic. If the deviation of an individual subject's value x_i from the group mean is defined by $x_i = X_i - \overline{X}$, then a short-circuit conjecture could suggest that the

$$\text{"Mean Deviation"} \quad MD = \frac{\sum x_i}{N}$$

could be the characteristic we are looking for. But don't forget the mutual counterbalancing of the left (negative) and right (positive) deviations around the mean to keep the scale "lever" in equilibrium. Therefore, $MD = 0$ for any data, since $\Sigma d_i = 0$ for any set of data. This is a good check that the mean has been calculated correctly. Thus, *MD* is not of any particular use as an index of variability. A solution could be to take the mean of the absolute value of the deviations to produce the

$$\text{"Mean Absolute Deviation"} \quad MAD = \frac{\sum |x_i|}{N}.$$

Although this statistical characteristic is not really "mad," it is not "cool" either. It is not used much despite the fact that originally it had been suggested by the great French mathematician Pierre-Simon Laplace (1749–1827). Instead, the problem has been solved by a famous British statistician of the 20th century, Sir Ronald Fisher the inventor of variance. Fisher's approach is not to take absolute values, but to avoid the negative values by computing their squares. Thus, instead of

MAD the better index of spread is the MSD, based upon a sum of squared deviations, a special case of the Bravais (1846) "product sum," or "sum of products."

$$\text{"Mean Squared Deviation"} \quad MSD = S_x^2 = \frac{\sum {x_i}^2}{N}.$$

It has become one of the most basic statistical concepts, and it is known under the name of variance. The variance of variable X is written out in longhand as:

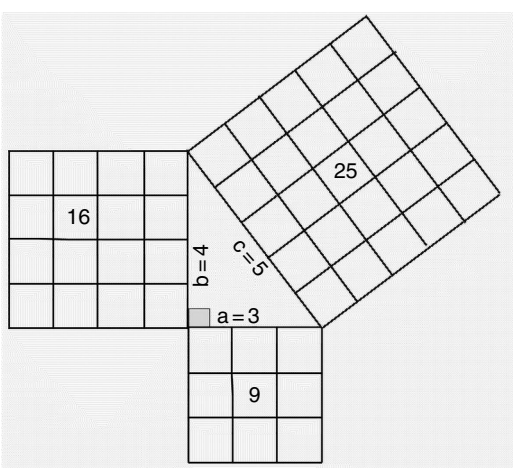
$$S_x^2 = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{N}$$

At the first glance, $MSD = S_x^2$ seems more complicated than *MAD*, since it is a number expressed in squares of the original measurement units. This last inconvenience can be quite intuitively corrected—after computing squared deviations and averaging them we can return to the original units by taking the square root of the variance to get the standard deviation, also called SD.

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{N}}$$

Galton first proposed the standard deviation in the late 1860s, just a few years after Bravais (1846) published the product sum. Note that *S*, by the arithmetic operations taken—and intuitively, too—is also a kind of "average" deviation like *MAD*.

Why was MAD dropped in favor of variance? At least two reasons can be given. One is that Gauss, the same mathematician who discovered the Gaussian distribution (the normal curve), and also the least squares principle on which regression is based, had the foresight to realize the profound advantage of using *squared* units for his least squares principle on which regression is based, which places it squarely in the mainstream of Euclidean geometry and the Pythagorean Theorem. Squared units are additive. Note in the classic 3–4–5 triangle at the right, that the sides are not additive ($3 + 4 \neq 5$), but their squares are ($9 + 16 = 25$). Sir Ronald Fisher made squared deviations the foundation of his variance statistic, published in 1918. It has become one of the foundation pieces of statistical theory. Regression and multiple regression (Appendices 4 and 5) also embody the additivity principle, in which the total variation in Y can be decomposed into predicted variation and error variation. Also, in CTT the total variance in a test is decomposable into true-score variance and error variance.



The second reason is that the standard deviation (SD), the square root of variance is a number, a statistic (that is, a sample index), with the special property of becoming[14] a parameter (a population index), which in some data distributions found in nature enables us to compute the relative frequency of the data from a normal curve table

14 In the first part of Section §13, in Table 1.3 and the accompanying text, we will explain the slight modification of the variance & standard deviation to transform them from sample statistics into unbiased population parameters.

and thus estimate probabilities of occurrence. As well, it is the parameter that allows us to estimate the attributes of a whole population if we know the SD in just a small random subsample. Thus, it makes it possible to establish the probable error of estimation for a whole population. Also, it enables us to compute the relative position of a subject within his or her sample on a standardized test, or even within a specific population. It also can be used to establish norms and standards of clinical normality, estimate the probability that a subject fulfills the norms, and so forth, as will be shown in Sections §12 and §13.

1.3.5 (§12) A Small Numerical Example, Finally

Didactic doctrine says that a theoretical presentation should be accompanied by examples that are illustrative and real. Didactic experience teaches that "real" examples are often too complicated to be illustrative. So, please accept the next artificial but simple example, patterned after the anthropomorphic measures of Galton, with the apology of the authors.[15] Imagine that we have selected $N = 6$ fathers and their sons from among the group of $N = 30$ father/son pairs whose data are shown in Figure 1.7. The head circumference values for the six selected fathers are shown in Table 1.1, together with the calculation of the mean of the head circumferences, the deviations of the circumferences from that mean, and the sum of squared deviations of the circumferences from the mean ($\sum x_i^2$).

The variance of the circumferences is then calculated by dividing this sum of squared deviations, 24, by N which gives 4 as the variance of the six fathers' head circumference measurements.

Table 1.1 Fictional head circumferences (X_i) of six fathers showing calculation of the sample total (T), the sample mean ($\overline{X}$), deviations from the mean (x_i), squared deviations from the mean, and the sum of the squared deviations from the mean ($\sum x_i^2$).

Family name	X_i	x_i	x_i^2
Abbott	26	3	9
Bailey	25	2	4
Cardon	23	0	0
Davis	22	−1	1
Evans	22	−1	1
Flanders	20	−3	9
$T = \sum x_i = 138$			$\sum x_i^2 = 24$
$\overline{X} = \frac{\sum X_i}{N} = \frac{138}{6} = 23$			

15 We have used small samples here to make the demonstration of the calculations for the descriptive statistics simpler and easier to follow. In actual practice, one would usually be interested in inferring from the sample to some population, so-called inferential statistics as introduced in the next section, and would expect to utilize substantially larger samples, selected at random, if possible, to ensure an adequate estimation of the distributional properties of the population.

$$S^2_{fathers} = \frac{\sum x_i^2}{N} = \frac{24}{6} = 4$$

Taking the square root of this, we get a standard deviation of 2, which is interpretable in inches. The sample standard deviation of the six fathers' head circumferences is therefore 2 inches.

$$S_{fathers} = \sqrt{4} = 2$$

Using the Excel command "=**normsdist(1)-.5**,"[16] it can be determined that, as shown in Figure 1.8, the proportion of total area under the normal curve expected to fall between the mean and one standard deviation unit above the mean is 0.3413, to four decimal places. Since the normal curve is symmetrical, the area of the normal curve between the mean and one standard deviation below the mean is also 0.3413. The sum of the two areas is 0.6826. In other words, if a data set is normally distributed, about 68% of the scores will fall between +1 and −1 standard deviations from the mean. The mean head circumference for this small sample of fathers is 23 inches, and we have just found the standard deviation to be 2 inches. So, we would expect about two-thirds of the scores to be located between 21 and 25 inches (23 inches ± 2 inches, that is, the mean plus or minus the standard deviation).[17]

With the command "=**normsdist(2)-.5**," it can also be seen that the area between the mean and two standard deviation units above the mean is 0.47725. Summing that area with a 0.47725 area from the left half of the distribution, the percent of scores falling between ± two standard deviation units from the mean is a little over 95% (95.45% to be more precise), so a *standard normal curve*, that

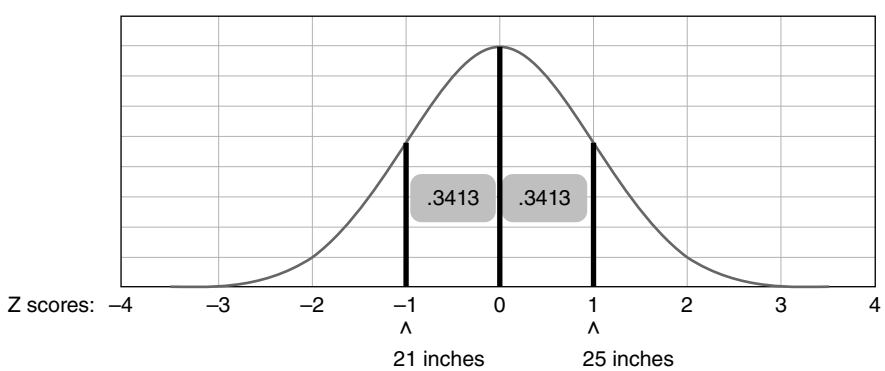


Figure 1.8 The area of the standard normal curve distribution between +1 and −1 standard deviation units is 0.3143 + 0.3413 = 6826, or about 68% of the area. Given that the six fathers have a mean head circumference of 23 inches with a standard deviation of 2, that means that we should expect about 2/3 of the fathers to be within this range. In Table 1.1, the middle four of the six fathers have circumference values of 22, 22, 23, and 25.

16 The Excel command "=normsdist(1)" requests the proportion of area under the normal curve that is below a *Z* score of 1.00, and the answer is 0.8413. But we want the area between a *Z* score of 1.00 and a *Z* score of 0.00 (the mean) so we append "−0.5" to the command to subtract the area in the lower half of the curve from 0.8413 to get 0.3413. For many years statistics books and psychometrics books have included statistical tables such as a table of areas under the standard normal distribution. Given the ready availability of statistical table functions such as this in spreadsheets such as Excel, the need for such tables is somewhat reduced.

17 With simple demonstration data such as this we would not expect the probabilities from the normal curve to be anything but very approximate. It is shown here to illustrate the process of using statistical tools and probability tables with simple data as a guide in preparation for use with research data.

Table 1.2 Fictional head circumferences (Y_i) of the sons of the six fathers in Table 1.1, showing the calculation of the sample total (T), the sample mean ($\overline{Y}$), deviations from the mean (y_i), squared deviations from the mean, and the sum of the squared deviations from the mean ($\sum y_i^2$).

Family name	Y_i	y_i	y_i^2
Abbott	26.4	2.95	8.703
Bailey	24.3	0.85	0.723
Cardon	23.5	0.05	0.003
Davis	24.4	0.95	0.903
Evans	19.6	−3.85	14.823
Flanders	22.5	−0.95	0.903
$T = \sum Y_i = 140.7$			$\sum y_i^2 = 26.055$
$\overline{Y} = \frac{\sum Y_i}{N} = \frac{140.7}{6} = 23.45$			

is, a normal curve with Z-score units for its horizontal axis, has approximately 68% of its area between one standard deviation above and below the mean and approximately 95% of its area between two standard deviations above and below the mean.

Consider now comparable measurements and calculations for the sons of these $N = 6$ fathers, for a comparison similar to that used by Galton. The head size measurements of the sons of these six fathers are presented in Table 1.2 together with their mean, their deviation scores from the mean, the squared deviation scores, and finally the sum of squared deviation scores. The variance and standard deviation are calculated for the head size measurements of the sons, using the same mathematical process just shown for the fathers' data. Using the sum of squared deviation scores of 26.055 from Table 1.2 the variance of the sons' head circumference values is calculated to be 4.34 and the standard deviation is the square root of 4.34, which is 2.08 inches:

$$S^2_{sons} = \frac{\sum y_i^2}{N} = \frac{26.055}{6} = 4.34$$

In Section §4 we discussed Galton comparing means and standard deviations of head circumferences for fathers and sons, and his development of the reversion or regression line to predict sons' head circumferences from that of their fathers. In this section, we have shown the statistical process of how those comparisons can be made. We have shown that not only do the sons have slightly larger mean head circumferences than their fathers for this small fictional example, but they also have a somewhat larger standard deviation. In the final section, Section §13, we will introduce inferential statistics in connection with the *Central Limit Theorem*, and demonstrate how statistical inference with sample sizes greater than 30 can give aggregated precision reminiscent of the double-dice tosses shown in Figure 1.6.

1.3.6 (§13) Does the Larger Head Circumference in the Sample of Sons Than Fathers Hold for the Whole Population? A Preliminary Sidestep to Statistical Inference

This question may be answered if we use a higher type of statistical reasoning, beyond *descriptive* statistics, that is, beyond those statistics that merely describe subjects within a particular sample. But this is possible only if some important statistical conditions have been met, enabling us to move

from descriptive statistics to *inferential statistics*. This branch of statistics offers methods for making an inference, a generalization, from a given representative *sample mean* and *sample standard deviation* to the corresponding characteristics of the whole population, but there is a caveat. Statistical inference works *only if the mathematical assumptions hold in reality*.

The first of these mathematical assumptions is a "randomly selected sample drawn from a clearly defined population." Obviously, a card register or a computer database of patients in a clinical psychologist's office is not comparable to a randomized representative sample of a city or a county. And, of course, the statistical meaning of the target "population" need not have a geographic or demographic reference at all. A *random sample* from a population is defined as *one in which each person or element in the population has an equal chance of being selected.*

An important reason why this topic is presented here and now, so early in the text, is that many students ask why standard deviation is sometimes symbolized by S or sometimes by σ, and why clinicians and researchers speak about "how many sigmas" from the mean is an acceptable limit of normality for a score. Moreover, some ask why the formula for calculating standard deviation is presented sometimes with division by N and elsewhere with division by $N-1$. The answers to questions of this kind require an understanding of inferential reasoning.

Suppose that we would be satisfied with just knowing how to compute a "one-valued" estimate of the true population mean or the true population variance, an estimate without additional information, say about its possible error. This is called a *point estimate*. In statistics, it is usual to use Greek letters to denote population characteristics, so we will use μ to represent a point estimate of the population mean and σ^2 to represent a point estimate of the population variance. From a statistical point of view, we can say that in a randomized sample, the mean $\overline{X}$ is the "best" *point estimate* of the population mean μ. Being the "best" point estimate requires that in a large series of many repeated random samples, the $\overline{X}$ will not systematically under-estimate nor over-estimate μ, the true population mean value. We can say, therefore, that the sample mean $\overline{X}$ is an *unbiased* estimator of the true population mean, μ.

But this unbiasedness does *not* hold for the descriptive sample variance S^2 ("s squared") as an estimate of the population variance σ^2 ("sigma squared"). The formula for the descriptive sample variance S^2 is given in Equation 1.1.

$$S^2 = \frac{\sum {x_i}^2}{N} = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{N} \qquad (\textit{Equation 1.1, descriptive sample variance})$$

It is nothing more than the average of the squared deviation scores, that is the sum of the squared deviations of scores from their mean ($\sum {x_i}^2$) divided by N. Since the standard deviation is the square root of variance, the unbiasedness also does not hold for the sample standard deviation.

If we were to use an ordinary descriptive sample variance such as this as the point estimate of the population variance, it would systematically underestimate the true population variance σ^2 by "one Nth" of its value. That is, if the true population variance were 12.0 and if the sample size were $N = 100$, the average of many such estimates would be about 11.88, underestimating the true population variance by 0.12, one one-hundredth of the value of σ^2. However, if the sample size were only 10, the average estimate of the variance over many trials would be about 10.80, underestimating the variance by 1.20, one-tenth of its true value. Obviously, estimation bias is much stronger with small sample sizes. The evidence for this sampling principle is presented at the end of this final section (Section §13) in Figure 1.15 and its accompanying text.

Fortunately, the demonstration given in Figure 1.15 not only verifies this principle, it also informs us as to what the magnitude of the systematic error is and how to correct for it. The correction is

simple. In the computation of the sample estimate of the population variance, instead of dividing the sum of squared deviations by the sample size N, we divide it by the *degrees of freedom* $N-1$,[18] as shown in Equation 1.2. The square root of this unbiased variance yields the unbiased standard deviation as shown in Equation 1.3. This is the correct formula to use for inferential estimation of the population standard deviation, the formula that will return an *unbiased point estimate of the true population standard deviation*.

$$\hat{\sigma}^2 = \frac{\sum {x_i}^2}{N-1} = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{N-1} \qquad (\textit{Equation 1.2, unbiased estimate of the population variance})$$

$$\hat{\sigma} = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{N-1}} \qquad (\textit{Equation 1.3, unbiased estimate of the population standard deviation})$$

Notice the *hat* (^) over the sigma in Equations 1.2 and 1.3. It is usual to place a "hat" over a statistical estimate to indicate that it is *statistically* "*good*," that is, "plausible" in one of the meanings used in statistical theory. Here, it happens to be a good inferential estimate, an unbiased estimate of a population characteristic on the basis of sample data. When the sample size is large, the degree of freedom adjustment makes little difference. But for small samples, the difference can be quite sizeable. We encounter another meaning of "good estimation" in the context of using a regression analysis to predict a Y variable from X variables, as shown in Appendices 4 and 5. Since the regression method fulfills certain requirements for statistical optimality, namely the "least squares" criterion, the hat ^ symbol is used for regression estimates as well.

Table 1.3 presents both the sample descriptive statistics and also the unbiased point estimates of the population parameters for the fictional sample data of six father-son pairs from Section §12. With the small N of only six fathers and six sons, this table reveals the somewhat sizeable disparity between the descriptive sample standard deviation *statistics* on the left and the unbiased point estimates of the population standard deviation *parameters*[19] on the right. Since the sample mean is an

Table 1.3 Descriptive sample standard deviations and means for the six father/son pairs from Tables 1.1 and 1.2 compared to the corresponding unbiased point estimates of the population values.

Sample descriptive statistics			**Unbiased point estimates of population parameters, based on recalculation of variances using degrees of freedom**		
	Means	**Standard deviations**		**Means**	**Standard deviations**
Fathers	$\overline{X} = 23.00$	$S_x = 2.00$	Fathers	$\overline{X} = 23.00$	$\hat{\sigma}_x = 2.19$
Sons	$\overline{Y} = 23.45$	$S_Y = 2.08$	Sons	$\overline{Y} = 23.45$	$\hat{\sigma}_y = 2.28$

18 In beginning to analyze a sample with $N = 6$ observations, all six values are free to vary during the calculation of the mean. The mean retains all 6 degrees of freedom. However, in the process of calculating a variance, the sum of squared deviations from the mean cannot be accomplished without first calculating the mean, which uses up one of the 6 degrees of freedom. Therefore, the number of degrees of freedom for calculating the variance is $N-1$, or 5 in this example.

19 Notice that a sample summary value, such as a mean $\overline{X}$, is referred to as a *statistic* and is expressed in ordinary Arabic letters, while the corresponding summary population value, such as the population mean μ, is referred to as a *parameter* and is expressed in Greek letters. One could say that statistic is to parameter as sample is to population.

unbiased estimate of the population mean, the means for the six fathers and the six sons on the left side of this table are the same as those on the right side. However, the standard deviations differ substantially. The computation of the standard deviations on the left follows the definition that variance is the arithmetic average of the squared deviation scores. On the other hand, for the unbiased point estimates of the population variances, on the right, it is essential to divide the sum of squared deviations by $N - 1$ rather than N. The square root of this unbiased variance is an important statistical achievement: an *unbiased point estimate of the unknown population standard deviation*. However, the point estimate could also be biased by inadequate sampling. Remember that it will only be an unbiased point estimate if the measurements were performed on a randomly selected sample.

A numerical index calculated from sample data in this statistically optimal way, by random sampling and by using the correct formula, is referred to as a valid and proper *statistical estimate*. There is one additional condition. It would be wise to obtain a sufficiently large sample to obtain a stable estimate of the mean. A small sample size was used here as part of a simplified and more transparent demonstration and to emphasize the strong effects of bias in a small sample. In an actual inferential estimate of population variance, a larger sample size would be called for. How large would depend on the purposes for the estimate and the level of precision needed.

If a larger sample of pairs of fathers and sons had been obtained from a properly randomized sample from a population, and if unbiased point estimates of the standard deviations of fathers and sons had been obtained in the manner described, then we could answer "yes" to the question in the title of this section. Under these conditions, it would be possible to determine whether the slightly larger head circumference for sons than fathers would hold for the whole population. Looking back to Section §11 where the logic of variance was explained, it should be emphasized that another reason for choosing the variance and the standard deviation in preference to the *MAD* statistical index as measures of spread is that no statistical theory is available for modifying *MAD* into a statistic that could enable optimal estimation of the corresponding population value in this way.

1.3.6.1 An Estimate of the Probable Error of a Point Estimate

As valuable as it is to have unbiased point estimates of population means, variances, and standard deviations, it is of substantially more value if it can also be combined with an estimate of the accuracy of those point estimates, that is, an estimate of probable error. This is true in psychometrics and also in the more general discipline of metrology, the scientific study of measurement. Brown (2021), in his article "Measuring measurement—What is metrology and why does it matter?," mentions the measurement of "uncertainty," another word for probable error, as one of the primary concerns of metrology. An evaluation of the uncertainty present in one's point estimate is essential to good measurement. To learn this process of estimating uncertainty, the probable error of one's measurements, it will be necessary to now take the "preliminary sidestep into inferential statistics" referred to in the subtitle of this section.

As a thought experiment, suppose we create a large collection of unbiased point estimates of the population mean taken from many random samples of a given population to evaluate how well they agree with one another. This collection of sample means is referred to as an *SDM*. The standard deviation of this SDM is called the *standard error of the mean* $\left(\hat{\sigma}_{\overline{X}}\right)$. It would indicate how large the spread is among all of the sample means. A large spread would be indicated by a large *standard error of the mean* $\left(\hat{\sigma}_{\overline{X}}\right)$, and a small spread would be reflected in a small standard error of the mean.

However, thanks to one of the principles of inference, the tedious task of assembling a large collection of samples from which to calculate $\hat{\sigma}_{\overline{X}}$ is not necessary. This short-cut principle is expressed

in Equation 1.4, which indicates that the variance of means is one *N*th as large as the variance of the corresponding raw scores.[20]

$$\hat{\sigma}^2_{\overline{X}} = \frac{\hat{\sigma}^2}{N} \qquad (\textit{Equation 1.4, the formula for the squared standard error of the mean})$$

This indicates that if one has an unbiased estimate of the population variance, Equation 1.4 can adequately estimate from that variance the *squared standard error of the mean*, or in other words the variance of the *SDM*. Taking the square root of this equation yields the *standard error of the mean* (Equation 1.5), which is the standard deviation of the SDM. This gives an unbiased estimate of the *probable error* for an estimated population mean.

$$\hat{\sigma}_{\overline{X}} = \frac{\hat{\sigma}}{\sqrt{N}} \qquad (\textit{Equation 1.5, the formula for the standard error of the mean})$$

For the fictional data in Table 1.3, the mean of the head circumferences for the six sons was found to be $\overline{Y} = 23.45$. The standard error for that mean is calculated using Equation 1.5.

$$\hat{\sigma}_{\overline{Y}} = \frac{\hat{\sigma}_y}{\sqrt{N}} = \frac{2.28}{\sqrt{6}} = \frac{2.28}{2.449} = 0.931 \qquad (\textit{Equation 1.5a}^{21})$$

This is an unacceptably large standard error of the mean, but that is because we have such a small N of 6 for this simple demonstration data. Suppose that the N' for the sample of these father/son pairs were 80 rather than 6. We now have a larger sample size for Y, denoted as Y', but having the same means and the same standard deviation as given in Table 1.3. The standard error for the mean of the sons is now of a more reasonable size.

$$\hat{\sigma}_{\overline{Y}'} = \frac{\hat{\sigma}_{y'}}{\sqrt{N'}} = \frac{2.28}{\sqrt{80}} = \frac{2.28}{8.944} = 0.255$$

To clarify the measurement uncertainty for this estimated mean as related to its standard error we will use a confidence interval. There are two steps in creating a confidence interval. The first is to decide on the confidence level chosen for the interval, and the second is to calculate the upper and lower score values that are the boundaries of that confidence interval. Perhaps we would like to be 95% confident that our population mean is within the confidence band. Using the Excel command "**=normsinv(.975)**,"[22] it is determined that a Z score of 1.96 has an area below it of 0.975 which corresponds to an area of 0.025 above it on the right side of the standard normal curve. Similarly, a Z score of −1.96 has an area of 0.025 below it on the left side. The 95% confidence region therefore lies

20 One would expect the variance of means to be less than the variance of raw scores given the effects of averaging, but the surprising thing is that the relationship is so simple. If one were to average raw data in randomly created groups of five, these means would be expected to have a variance that is one-fifth that of the raw scores. If one averages in groups of ten, the variance of the means would be one-tenth of the variance of raw scores. In the language of statistical inference this quantity is referred as the *variance of the sampling distribution of means*, but it is also called, and is equivalent to, the *squared standard error of the mean*. This equation is also an aspect of the *Central Limit Theorem* introduced in Section §7. A demonstration of its validity is found in Figure 1.14 and accompanying text at the end of this section.

21 Notice that this is still Equation 1.5, but now applied to variable *Y*. When introducing the formula, it was shown for variable *X*. In other words, in its first appearance the symbol for the formula was $\hat{\sigma}_{\overline{X}}$, which is read as "the standard error of the $\overline{X}$ mean." In the second appearance the symbol was $\hat{\sigma}_{\overline{Y}}$, which is read as "the standard error of the $\overline{Y}$ mean"—the same formula, but applied to a different variable.

22 The Excel command "normsinv(.975)" requests the Z score value that has an area below it of 0.975, which returns a Z value of 1.96.

between $Z = -1.96$ and $Z = 1.96$ on the standard normal curve. The confidence interval now needs to be expressed in raw score units rather than Z score units. Equation 1.6 will create the desired confidence interval.

$$CI_{95} = \overline{X} \pm Z_{0.025}\,\hat{\sigma}_{\overline{X}} \qquad (\textit{Equation 1.6, confidence interval for a single mean})$$

Applying this equation to variable Y' with a sample size of 80 and the same mean and standard deviation given for the sons in Table 1.3, we obtain a confidence interval of 23.45 ± 0.50.

$$CI_{95} = \overline{Y} \pm Z_{0.025} \quad \hat{\sigma}_{\overline{Y}} = 23.45 \pm (1.96)(0.255) = 23.45 \pm 0.50$$

The lower limit of this confidence interval is equal to $23.45 - 0.50 = 22.95$. The upper limit is $23.45 + 0.50 = 23.95$. The 95% confidence interval for the population mean of the head circumferences of sons therefore extends from 22.95 to 23.45. Since the mean head circumference of fathers was found to be 23.00, and lies within the confidence band for sons, we conclude that we do not have sufficient evidence to say that the larger head circumference in the sample of sons than fathers holds for the whole population.

In the past several pages, some of the principles of statistical inference have been introduced: how to obtain a point estimate of the population mean, how to obtain an unbiased estimate of the population variance that can be used to derive the standard error (that is the probable error) of that population mean, and how to combine these two statistics to construct a descriptive and interpretable confidence interval within which one can be 95% confident the true population mean lies. These are some of the uses and principles of stochastic (probabilistic) modeling, one of the three preconditions specified in the last four paragraphs of Section §5 for creating "bona fide research data".

In the following three subsections, additional examples of stochastic models will be presented to summarize and clarify the central fundamental principles of statistical inference. These stochastic models will function like brief "parables" that demonstrate principles in a mathematically simpler form using binary data. In the fourth and final remaining subsection, we will expand from binary to multinary data, using the multinomial distribution, to illustrate the principles of inference applied to interval/ratio data.

1.3.6.2 Classical Probability Theory and the Binomial Distribution: Creating a Simple Stochastic Model

Statistical inference is based upon classical probability theory, the mathematical creation of a coterie of "natural philosophers" during the Age of Enlightenment. It began with the pioneering work of Jakob Bernouli in the 17th century and extended into the 18th and 19th centuries with the work of theorists such as Pierre-Simon Laplace (Daston, 1995). Though "probability theory" sounds daunting, the fundamental principles can be made quite understandable, as stated by Laplace in the introduction to his 1812 book, *Théorie Analytique des Probabilités*,

> The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which ofttimes they are unable to account.

Classical probability theory and inferential statistics are often taught using binary data, which serves as a simpler proxy for quantitative data. This works well. The mathematical procedures for means and variances are often the same for binary data as for fully quantitative interval/ratio data. For an illustrative example, suppose that surveys have been taken of citizen approval of a new

law adopted in a small town. One survey measures the level of approval using a 10-point scale. A second survey is conducted using binary "yes" or "no" answers. The first six persons to answer the quantitative survey give ratings of 5, 7, 3, 4, and 2. On the other hand, the first six persons to answer the binary second survey respond with "yes," "yes," "no," "yes," and "no," which responses are then "dummy coded" as 1, 1, 0, 1, 0. The mean of the quantitative survey is calculated as:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{5+7+3+4+2}{5} = \frac{21}{5} = 4.2$$

The mean of the second survey, which is actually a proportion since it was calculated for binary data, is found to be:

$$p = \overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{1+1+0+1+0}{5} = \frac{3}{5} = 0.60$$

In both cases, the response values for the six persons are just summed and divided by N, the number of responses. Three out of the five responses on the second survey, or 60% of them, were "yes" responses. This corresponds to a proportion of 0.60. It is mathematically equivalent to the mean calculated in the usual way. It is the mean of a variable consisting of only ones and zeros. Simply stated, the proportion is a *special case of the mean*. It is the mean of a binary variable.

The descriptive sample variance of the quantitative survey is calculated as:

$$S^2_{quant} = \frac{\sum x_i^2}{N} = \frac{\sum \left(X_i - \overline{X}\right)^2}{N} = \frac{(0.8)^2 + (2.8)^2 + (-1.2)^2 + (-0.2)^2 + (-2.2)^2}{5} = \frac{14.8}{5} = 2.96$$

The descriptive sample variance of the second survey, the binary one, can be calculated in the same manner, simply the sum of the squared deviations of scores from their mean, divided by N:

$$S^2_{binary} = \frac{\sum x_i^2}{N} = \frac{\sum \left(X_i - \overline{X}\right)^2}{N} = \frac{(0.4)^2 + (0.4)^2 + (-0.6)^2 + (0.4)^2 + (-0.6)^2}{5} = \frac{1.2}{5} = 0.24$$

There is also, however, a much simpler way to calculate the variance of the binary variable. We just calculated the proportion for this binary variable, that is its mean as $p = \overline{X}$, and found it to be 0.60. If q is defined as "one minus p," the variance of the binary variable may be calculated simply as pq, the product of p and q as shown in Equation 1.7.

$$S^2_{binary} = pq = p(1-p) = (0.6)(0.4) = 0.24 \qquad (\textit{Equation 1.7, the variance of a proportion})$$

Surprisingly, this variance is not only the descriptive sample variance of the binary data of the second survey, it is also the unbiased point estimate of the corresponding variance of the binary data in the population.[23] The square root of this variance is therefore the unbiased standard deviation of the binary data in the population. In other words, the formula $\hat{\sigma} = \sqrt{pq}$ is used to obtain an *unbiased point estimate of the true population standard deviation of binary data*, which for our example data from the yes/no survey is found to be: $\hat{\sigma} = \sqrt{pq} = \sqrt{(0.6)(0.4)} = \sqrt{0.24} = 0.490$.

23 With data that is continuous and quantitative, the variance can only be calculated after the mean has first been calculated, since the mean is first needed to obtain the deviations from the mean for the "sum of squared deviations" used in calculating the variance. This uses up one degree of freedom. However, as demonstrated above, for binary data the variance can be calculated directly from the proportion without obtaining squared deviations from the mean, so there is no loss of degrees of freedom. It follows that, for binary data, the formula $\hat{\sigma}^2 = pq = p(1-p)$ gives an unbiased point estimate of the population variance, and $\hat{\sigma} = \sqrt{pq} = \sqrt{1-p}$ gives an unbiased point estimate of the true population standard deviation.

This estimate of the standard deviation of the binary data in the population can in turn be used to obtain an estimate of probable error, that is, of the *standard error of the proportion*, by dividing *pq* by *N* before taking the square root, as shown in Equation 1.8.

$$\hat{\sigma}_p = \sqrt{\frac{pq}{N}} \qquad (Equation\ 1.8,\ standard\ error\ of\ the\ proportion)$$

For this small example, the *standard error of the proportion* is calculated to be:

$$\hat{\sigma}_p = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(0.6)(0.4)}{N}} = \sqrt{\frac{0.24}{5}} = \sqrt{0.048} = 0.219$$

Additional fundamental principles of probability theory will now be demonstrated using the *binomial distribution* to create sampling distributions of means (or more precisely of proportions) for binary data. Consider, for example, the common act of coin tossing for our first stochastic model. Coin tossing has the advantage of being familiar, simple, concrete, and memorable.

The Multiplication Rule for Independent Joint Events If a fair coin[24] is tossed three times, one of eight things can happen. It can come up with three heads in a row (HHH) as shown in the first branch of Figure 1.9. Or, it could come up with two heads in one of three ways (in the order of HHT as shown in the second branch, or HTH as in the third, or THH as in the fifth). Or, it could come up with a single head in one of three ways (in the order HTT as shown in the fourth branch, THT as in the sixth, or TTH as in the seventh). Or, finally, it could come up with zero heads in one way (TTT), as shown in the eighth branch of Figure 1.9. To the right of the branching diagram is a pattern listing in the same order from top to bottom as the branching diagram, but with probability calculations

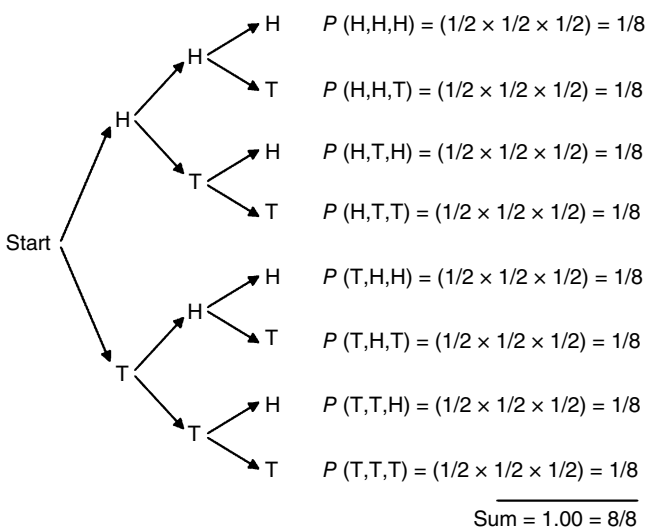


Figure 1.9 Branching diagram showing the eight possible patterns of heads and tails created by "fair coin" tossing ($p = 1/2$) in these-toss trails, and the calculation of each their respective probabilities.

24 A "fair coin" is a coin that comes up heads half of the time and tails half of the time. That is, over the long haul the expected value of its proportion of times coming up heads, its mean number of heads is 0.50, when "heads" is dummy coded as "1" and tails is dummy coded as "0."

added for each of the eight outcome branches. Since this is a fair coin, the probability for each head is always ½ and the probability for each tail is always ½ also. The probability for each branch of the diagram is just the product of individual probabilities for each of the three events (H or T) in sequence, which comes out to 1/8 in all eight patterns of the $p = ½$ model. This is known as *the multiplication rule for independent*[25] *joint events*, Equation 1.9. It is used for "AND" probabilities. It is written as:

$$P(A \text{ AND } B) = P(A) \times P(B) \qquad (\textit{Equation 1.9, the multiplication rule})$$

In set theory notation, "AND" probabilities are referred to as "intersection" probabilities. The multiplication rule is written in set notation as "the probability of A intersect B is equal to the probability of A multiplied by the probability of B":

$$P(A \cap B) = P(A) \times P(B)$$

The multiplication rule can, of course, also be generalized to three or more independent joint events as shown in this set-theoretic version of the equation:

$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C) \qquad (\textit{Equation 1.9 expanded for three joint events}^{26})$$

As Laplace stated, probabilities are nothing but "common sense reduced to calculus," or in other words common sense logic expressed with numerical calculations. From a logical standpoint, the probability of getting a single head in one toss is one-half (½), but the probability of getting two heads in two tosses is half of that half, or one-fourth ($½ \times ½ = ¼$), as can be seen by examining the last two tosses in the first four outcome patterns of Figure 1.9 (HH, HT, TH, and TT, only one of which is two heads). The probability of getting three heads in three tosses is half-of-a-half-of-a-half, or one-eighth ($½ \times ½ \times ½ = ⅛$). Of course, since this example in Figure 1.9 gives results for tossing a fair coin, the probability is always 0.50 whether heads or tails, and the probability of TTH or any other one of the eight outcome patterns is always one eighth ($½ \times ½ \times ½ = ⅛$) as shown in the individual calculations of Figure 1.9. This is the simple and unique branching diagram that results when $p = q$, and each of them is therefore equal to ½ (0.50).

It must also be noted that this is a probability model, so the outcomes are *mutually exclusive*, meaning that if the outcome is HHT it cannot also be THT or any other one of the remaining six specified outcomes of the eight branches. The outcomes are also *exhaustive*, meaning there are no other possibilities than the eight that are listed. Finally, the *probabilities must sum to 1.00*, which, of course, they do since there are eight possible outcomes specified, and each has a probability of one-eighth as shown in the right half of Figure 1.9. These two paragraphs are an example of the "common sense" part of Laplace's statement, the logic of probability. The following

25 Some joint events are *dependent joint events*, meaning that the first event occurring has an impact on the probability of the second event occurring. For our purposes, we will only consider independent joint events, that is joint events where the occurrence of the first event has no effect on the probability of the second event. The equation for the *general multiplication rule of probabilities* is $P(A\_\text{AND}\_B) = P(A) \times P(B|A)$, where $P(B|A)$ refers to "the probability of B given that A has occurred," in other words $P(B)$ can be changed by the prior occurrence of event A. For more information about the general rule, go to Khan Academy online or do a web search on "the general multiplication rule of probabilities."

26 When we ask the question "how many possible patterns are there in three (or four, or five, etc.) tosses of a coin?" we are thinking of this as a "joint event" consisting of three, or four, or any number of combined single events. Equation 1.9 can therefore be expanded to any number of independent joint events in the same way we expanded it from two-toss joint events to three-toss joint events.

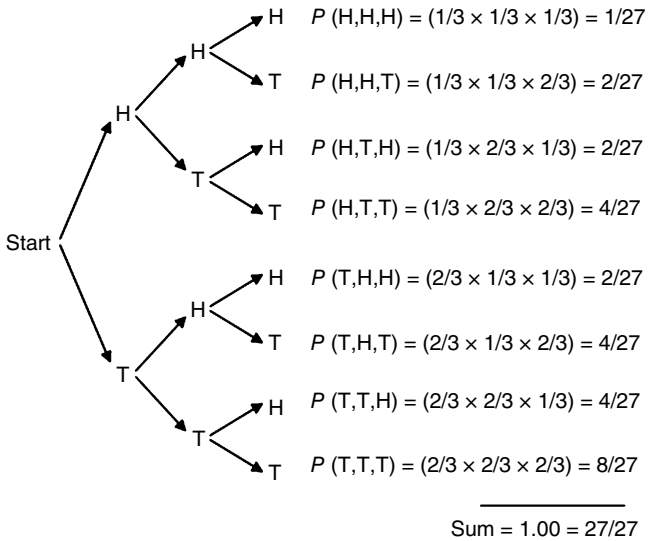


Figure 1.10 Branching diagram showing the eight possible patterns of heads and tails created by "weighted coin" tossing (p = 1/3) in three-toss trails, and the calculation of each of their respective probabilities.

paragraphs and their accompanying figures and tables spell out the "reducing-it-to-calculus" part, that is expressing the logic as calculations.

Whereas the probability model in Figure 1.9 has equal coin-tossing probabilities for heads and tails, our second stochastic model in Figure 1.10 is more complex. It is a "weighted coin" model with probabilities of $p = ⅓ = 0.333$ and $q = ⅔ = 0.667$. Although these two stochastic models have similar structures, each is a separate and distinct probability model, because they have differing probabilities. Whereas the eight branches in Figure 1.9 "fair coin" branching diagram have the same probability (all are $P = 1/8$), those in Figure 1.10, the "weighted coin" model do not. Because it is more complex, the "weighted coin" branching diagram gives more insight into the nature of probability.

The Addition Rule for Mutually Exclusive Composite Events The probability for each branch of Figure 1.10 can be obtained using Equation 1.9. For example, the probability for the second branch down from the top (HHT) is calculated by defining $P(A) = 1/3$ in the equation as "the probability of H, a head, on the first toss," $P(B) = 1/3$ as "the probability of H on the second toss," and $P(C) = 2/3$ is defined as "the probability of T, a tail, on the third toss." As shown in Figure 1.10, the joint probability of this HHT sequence in the second branch is calculated to be the product of those three probabilities, which equals 2/27.

$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C) = P(H) \times P(H) \times P(T) = \left(\frac{1}{3}\right) \times \left(\frac{1}{3}\right) \times \left(\frac{2}{3}\right) = \frac{2}{27}$$

These same three probability values also appear (although with their order permuted) in the other two outcome patterns that are characterized by the description of "getting two heads in three tosses of a weighted coin," namely HTH, the third branch down; and THH in the fifth branch. It follows that they each also have a probability of 2/27. If one asks the question "what is the probability in the model of Figure 1.10 of getting exactly two heads in three tosses of a coin?," the answer must be "the combined sum of the probabilities of these three patterns, HHT, HTH, and THH,"

since all three fit this same description of "two heads in three tosses of a coin," and none of the patterns in the other five remaining branches does.

$$P(2 \text{ heads}) = P(\text{HHT}) + P(\text{HTH}) + P(\text{THH}) = \frac{2}{27} + \frac{2}{27} + \frac{2}{27} = 6/27$$

This is *the addition rule of mutually exclusive*[27] *composite events*.[28] It is used for "OR" probabilities, the probabilities that are additive. Equation 1.10 below gives the basic equation for the addition rule with two mutually exclusive events and Equation 1.10a is an expanded version for three such events.

$$P(A \text{ OR } B) = P(A) + P(B) \quad (\textit{Equation 1.10, the addition rule for mutually exclusive events})$$

In set theory notation, "OR" probabilities are referred to as "union" probabilities. The basic addition rule equation for two mutually exclusive composite events is written in set notation as:

$$P(A \cup B) = P(A) + P(B)$$

The addition rule can also be generalized to three mutually exclusive composite events as shown in the following set theory version of the equation

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

(*Equation 1.10a expanded for three composite events*)

Besides the additive composite probability of the three mutually exclusive joint events for "two heads in three tosses" just demonstrated, there is also an additive composite probability for three joint events that share the description of "one head in three tosses." The first of these joint events is the one for HTT, the fourth branch down in Figure 1.10. It is found to have a probability of 4/27 (Equation 1.10a).

$$P(\text{HTT}) = P(\text{H}) \times P(\text{T}) \times P(\text{T}) = \left(\frac{1}{3}\right) \times \left(\frac{2}{3}\right) \times \left(\frac{2}{3}\right) = \frac{4}{27}$$

The other two are the sixth and the seventh branches down from the top in Figure 1.10. They also each have a probability of 4/27.

$$P(\text{THT}) = P(\text{T}) \times P(\text{H}) \times P(\text{T}) = \left(\frac{2}{3}\right) \times \left(\frac{1}{3}\right) \times \left(\frac{2}{3}\right) = \frac{4}{27}$$

$$P(\text{TTH}) = P(\text{T}) \times P(\text{T}) \times P(\text{H}) = \left(\frac{2}{3}\right) \times \left(\frac{2}{3}\right) \times \left(\frac{1}{3}\right) = \frac{4}{27}$$

27 Mutually exclusive means that if one occurs the other cannot. Heads and tails are mutually exclusive. If a coin toss is "heads" it cannot also be "tails." If a three-toss joint event yields "3 heads" it cannot also yield "2 heads," since the "OR" is "exclusive or," exactly 2 or exactly 3. For our purposes, we will only consider mutually exclusive composite events. The equation for the general case addition ruleis: $P(A\_\text{OR}\_B) = P(A) + P(B) - P(AandB)$. here $P(AandB)$ is the probability of an event that qualifies as both event A and also as event B. Dropping this subtractive third term of the general case addition rule formula yields the formula for mutually exclusive events.

Consider a small sample of five digits "1, 2, 3, 4, 5." Define event A as "a digit greater than 2" and event B as "an even digit." With only these five digits to choose to draw from the probability of event A is $P(A) = 3/5$, since three of the five digits are greater than 2. The probability of event B is $P(B) = 2/5$ (since 2 and 4 are both even digits). One of the digits, 4, is both an even digit (event A) and also a digit greater than 2 (event B), so it might be counted twice. The general addition rule corrects for this: $P(A\_\text{OR}\_B) = P(A) + P(B) - P(AandB) = 3/5 + 2/5 - 1/5 = 4/5$. The special case formula for "mutually exclusive events" fails to correct for it: $P(A\_\text{OR}\_B) = P(A) + P(B) = 3/5 + 2/5 = 5/5$. It counts the 4 in both event A and event B without subtracting out the double count. The digit 1 doesn't fit either event, but the other four digits fit A only (3, 5) or B only (2), or both (4).

28 They are composite events in that each has a unique composition, a unique pattern of outcomes.

The additive composite probability of the three independent joint events for "one head in three tosses" is found also using Equation 1.10a.

$$P(1 \text{ head}) = P(A \cup B \cup C) = P(\text{HTT}) + P(\text{THT}) + P(\text{TTH}) = \frac{4}{27} + \frac{4}{27} + \frac{4}{27} = 12/27$$

The two single outcome branches of the eight in the branching diagram of Figure 1.10 are the first branch which has three heads in three tosses, and the eighth branch which has zero heads in three tosses.

$$P(\text{HHH}) = P(\text{H}) \times P(\text{H}) \times P(\text{H}) = \left(\frac{1}{3}\right) \times \left(\frac{1}{3}\right) \times \left(\frac{1}{3}\right) = \frac{1}{27}$$

$$P(\text{TTT}) = P(\text{T}) \times P(\text{T}) \times P(\text{T}) = \left(\frac{2}{3}\right) \times \left(\frac{2}{3}\right) \times \left(\frac{2}{3}\right) = \frac{8}{27}$$

We can now create an additive composite of all eight branches of the branching diagram of Figure 1.10, and we see that it is indeed a probability model with the sum of the probabilities being equal to 1.00.

$$P(\text{All 8 branches}) = P(\text{HHH}) + P(2 \text{ heads}) + P(1 \text{ head}) + P(\text{TTT})$$
$$= \frac{1}{27} + \frac{6}{27} + \frac{12}{27} + \frac{8}{27} = 27/27$$

We have just created a stochastic model of what is referred to as the *binomial distribution*.

1.3.6.3 The Binomial Distribution as a Simple and Clear Example of a Sampling Distribution of Proportions

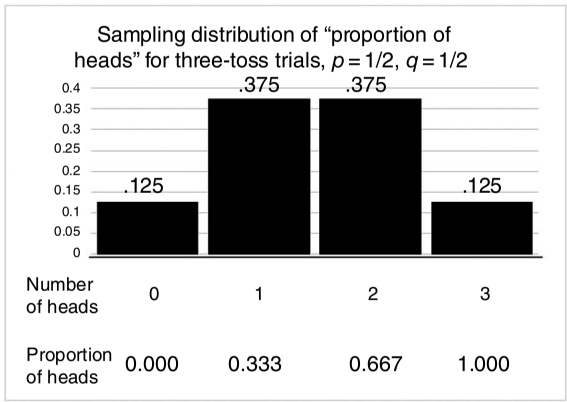
From the probabilities of the eight possible *independent joint events* in Figures 1.9 and 1.10, we can now create a sampling distribution of proportions for each of these two stochastic models. Although there are eight possible joint events for each, that is eight "branches" or patterns of results, there are only four possible outcomes in terms of the number of heads from these eight patterns for each model. The four outcomes are three heads, two heads, one head, and zero heads. The corresponding "mean number of heads" (that is, proportions) are 1.000 for the three heads outcome (HHH), $p = 0.667$ for the two heads outcome (HHT, THH, or THH), $p = 0.333$ for the one head outcome (HTT, THT, or TTH), and $p = 0.000$ for the no heads outcome (TTT).

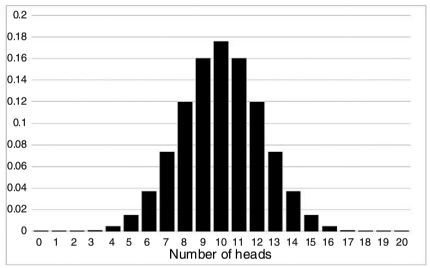
The probability model for each of these two stochastic models, one for the fair coin data and the other for the weighted coin data, can now be completed by creating a summary *sampling distribution of proportions* for each as shown in the two left panels of Figure 1.11. These two sampling distributions are an example of, and nothing less than an SDM, where the means are proportions, that is, the means of binary data. Each is a summary distribution of all the means that are possible for a given sample size, $N = 3$ tosses in this case, in sampling from a particular population. Taken together these identify the relative probability of each outcome.

It is unfortunate that the four "proportions" (the mean-number-of-heads values for each model) and the "probabilities" (the stochastic likelihood of the outcome of each) both start with the letter "*p*," making things a little confusing. The two panels, upper and lower, on the left side of Figure 1.11 provide a tabular comparison of the outcome probabilities of the fair coin model and the weighted coin model. The two bar graph panels in the middle of Figure 1.11, the fair coin one above and the weighted coin one below, provide a comparison of the shape of these two sampling distributions of the means of binary variables. The graphical distribution for the $N = 3$ fair coin model is obviously symmetrical but the one for the weighted coin model is clearly right-skewed. A small sample size of $N = 3$ was used to make the calculations simple and clear. But now, for a good graphical comparison

Sampling distribution of the "proportion of heads" for three-toss trials of a fair coin, with $p = 1/2$ and $q = 1/2$

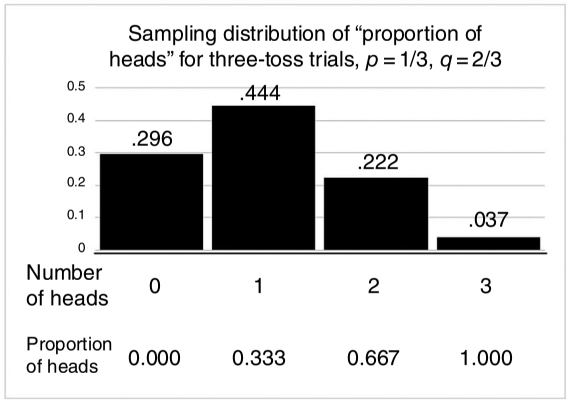
Number of heads	Proportion of heads	Probabilities (p)
3	1.000	p = 1/8 = .125
2	0.667	p = 3/8 = .375
1	0.333	p = 3/8 = .375
0	0.000	p = 1/8 = .125
		Sum = 1.000





Sampling distribution of the "proportion of heads" for three-toss trials of a weighted coin, with $p = 1/3$ and $q = 2/3$

Number of heads	Proportion of heads	Probabilities (p)
3	1.000	p = 1/27 = .037
2	0.667	p = 6/27 = .222
1	0.333	p = 12/27 = .444
0	0.000	p = 8/27 = .296
		Sum = 1.000



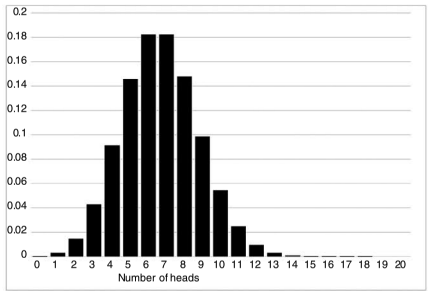


Figure 1.11 Summary tables and figures displaying the pattern of the "fair coin" results from Figure 1.9 appear in the upper three panels of Figure 1.11, with tables and figures of the "weighted coin" results from Figure 1.10 in the lower three panels. Tables of the sampling distribution of proportion of heads for the "fair coin" model compared to the "weighted coin" model are presented in the left panels of Figure 15, with a comparison of their respective bar graphs in the two middle panels of that figure. On the far right are bar graph representations for these two models applied to 20-toss trails, with the "fair coin" model on top and the skewed "weighted coin" model below.

of the sampling distributions, it would be good to have a larger sample. We therefore repeat these same calculations using a sample size of $N = 20$ both for the fair coin model and also for the weighted coin model, and obtain the resultant bar graphs shown in the two right panels of the figure. We expect the bar graph for the fair coin model to become more and more Gaussian as the sample size increases. It does. The $N = 20$ fair coin model graph in the top right panel of the figure is clearly bell-shaped. One could even say that, although it is a discrete distribution rather than a continuous curve, it approaches a proto-Gaussian shape. Interestingly, the corresponding $N = 20$ weighted coin graph in the bottom right panel of Figure 1.11 is also reasonably bell-shaped. The right skew of the probability model shows up primarily in the offset of the midpoint of the curve to the left side of the *X*-axis of the curve, centering on the $p = ⅓$ location in contrast to the $p = ½$ location for the fair coin graph.

The two-bar graph comparison on the right of Figure 1.11 is a good example of the *Central Limit Theorem*. It illustrates that regardless of the shape of the population distribution, the SDM from that population distribution will approach Gaussian shape as sample size *N* increases. The population for the weighted coin model is skewed with a 2/3 probability for tails, and a 1/3 probability for heads, but the sampling distribution of proportions is reasonably bell-shaped even with a sample size of only $N = 20$.

This section of Chapter 1 has a subtitle of "...a preliminary sidestep into statistical inference." Figure 1.12 continues that sidestep with a one-figure summary of the essential logic of statistical inference and sampling theory. Every statistical inference question involves data distributions of three kinds: (1) one or more samples, (2) the population distribution from which those samples are drawn, and (3) the sampling distribution of a particular summary statistic such as a mean or a proportion. Since a proportion is a special case of a mean, the process used for demonstrating statistical inference with proportions can also be applied to sampling distributions of means of fully quantitative data. Of these three types of distribution, the first one entitled "one or more samples" is real, that is, it is observed empirical data. The other two, the population and the SDM/proportions are hypothetical and mathematically constructed. Notice in this figure that example sample spaces are shown for three sizes of samples: a sample size of $N = 3$, a sample size of $N = 4$, and a sample size of 120. This figure summarizes the "big picture" of the logic of creating sampling distributions of proportions or of other types of means.

Examining the hypothetical sample space of the eight possible samples for an $N = 3$ binomial you will recognize it as a summary of what was shown in the branching diagram of Figure 1.10. The first of the eight icons in this sample space, the one at the top, has three heads on the right and zero tails on the left. Right below that icon is a set of three similar icons, identical to one another, each with two heads on the right and one tail on the left (but with each of the three representing one of the three different permuted orders of HHT, HTH, or THH[29]). Below that is another set of three icons. These have one head on the right and two tails on the left (HHT, THT, TTH). At the bottom is an icon with three tails on the left and zero heads. Below that sample space is its corresponding bar graph showing the shape of the SDM for this $N = 3$ binomial.

A similar sample space is shown to the right of that one, but for a sample size of $N = 4$. This sample space has sixteen possible sample outcomes. The first of the icons for this sample space, the one at the top, has four heads and no tails. Right below that is a set of four icons, each having three heads on the right and one tail on the left. Below this is six icons, each having two heads on the right and two tails on the left. Below this is another set of four icons with one head on the right and three tails on the left, and at the bottom is a single icon with four tails on the left and zero heads on the right. This $N = 4$ sample space also has its corresponding *SDM* summary bar graph below it. Each of the

29 As shown in Figure 1.10, these orders reflect the tosses (first, second, or third) on which heads or tails appear.

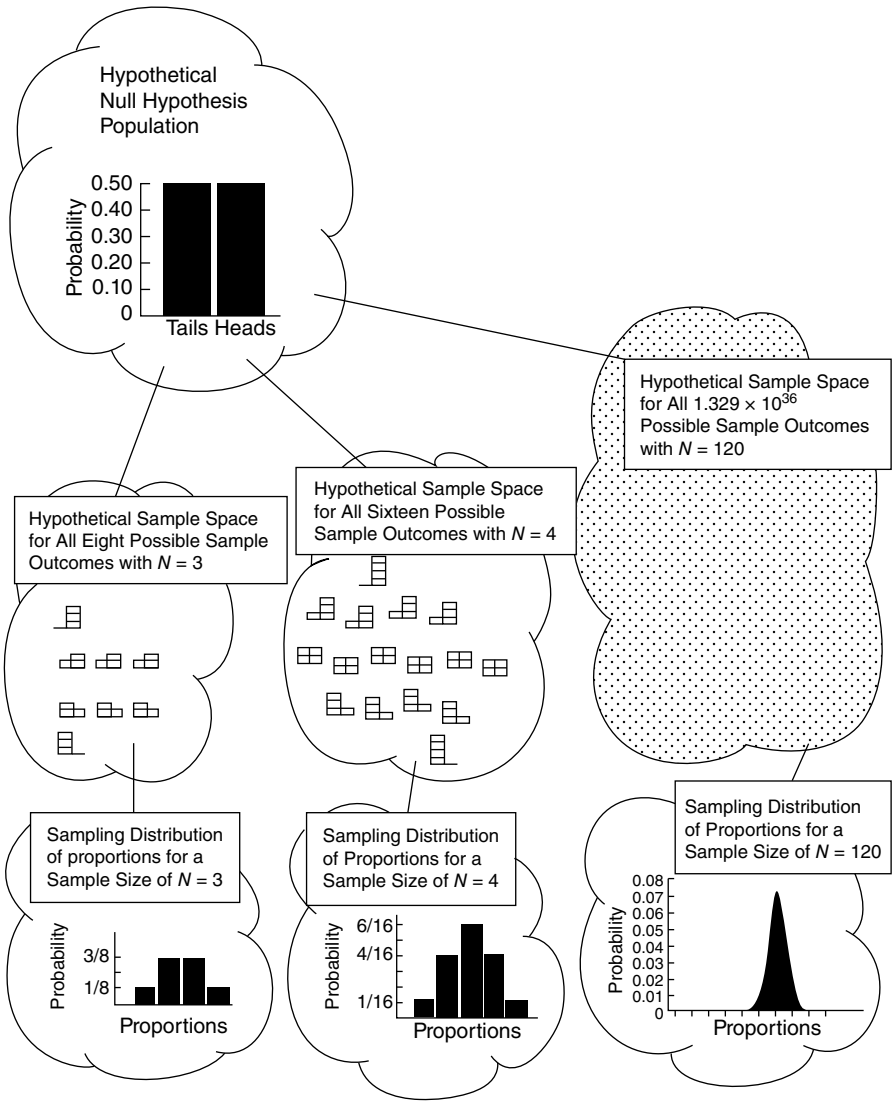


Figure 1.12 Single figure summary of the fundamental principle of statistical inference showing the three kinds of distributions: (1) a population, (2) samples from that population, and (3) sampling distributions of proportions from the sample spaces. *Source:* Reproduced from page 55 of Brown, Hendrix, Hedges, and Smith (2012).

bars in these bar graphs of the sampling distribution of means shows the relative frequencies (probabilities) of each binary mean (proportion) value.

1.3.6.4 Elaboration of Higher Level Calculations for the Binomial Distribution

The process of "expanding the binomial" was pursued early in the history of mathematics. To expand an expression such as $(p + q)^n$ means simply to multiply it out. Students in an elementary algebra class, for example, may learn to expand $(p + q)^2$ as $(p + q)^2 = p^2 + 2pq + q^2$. One can first multiply the $(p + q)$ binomial by p to obtain $(p^2 + pq)$. Next, multiply $(p + q)$ by q to obtain $(pq + q^2)$. Finally, sum the two conformable terms to obtain $p^2 + pq + pq + q^2 = p^2 + 2pq + q^2$.

This same process can be followed to obtain $(p + q)^3$ by multiplying $p^2 + 2pq + q^2$, first by p and then by q, and finally combining the two pairs of conformable terms to obtain the following equation.

$$(p + q)^3 = p^3 + 2p^2q + 2pq^2 + q^3$$

An astute French mathematician, Blaise Pascal (1623–1662)[30] noticed a pattern in these binomial expansions for various powers of $p + q$. When the expressions for $p + q$ taken to the first, second, and third power are placed above one another in a vertical array, with all the multiplicative coefficients showing, the pattern can be seen.

$(p + q)$ to the first power is $1p + 1q$

$(p + q)^2$ to the second power is $1p^2 + 2pq + 1q^2$

$(p + q)^3$ to the third power is $1p^3 + 3p^2q + 3pq^2 + 1q^3$

The coefficients for the algebraic expansion of the binomial for $(p + q)^4$ are shown in the Pascal's triangle of Figure 1.13 (which will be explained in the following paragraphs). As can be seen from the expansion of the binomial for $(p + q)^3$ and for $(p + q)^2$ all that is needed for the expansion of the binomial besides the coefficients is the appropriate powers of p and q as multipliers. Examining the "three-toss" binomial for $N = 3$ shown earlier, it is apparent that the appropriate powers of p start at N and decrease by 1 at each step from left to right, that is, from p taken to the third power (p^3) down

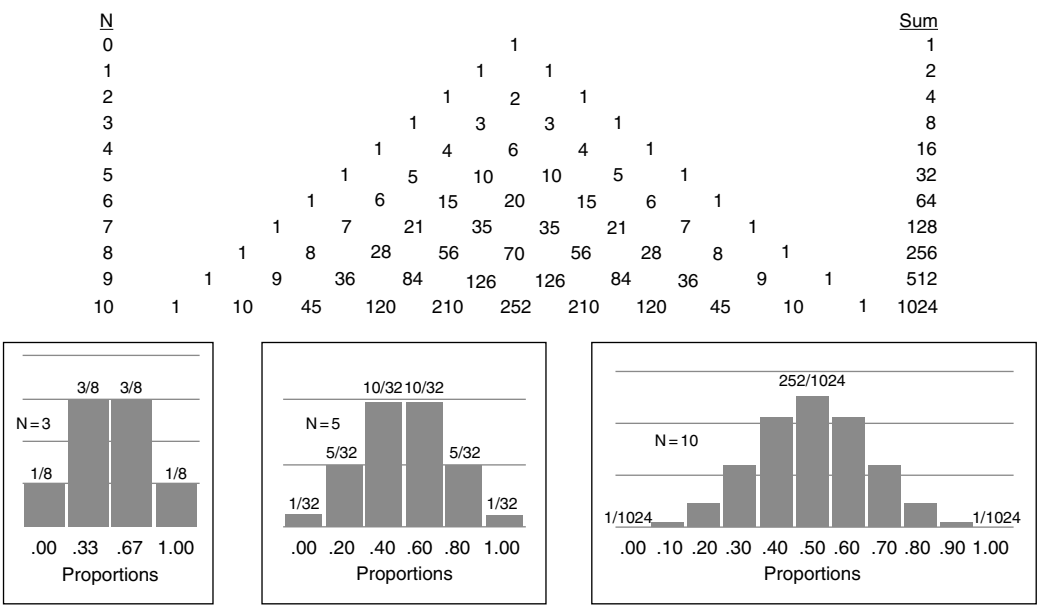


Figure 1.13 Pascal's triangle coefficients for all coin tosses beginning with $N = 1$ coin and extending to $N = 10$ coins. Also, three sampling distributions of proportions were created from these coefficients for samples of size $N = 3$, $N = 5$, and $N = 10$, with each having $p = 1/2$.

30 This had been noticed by several mathematicians in earlier centuries, but Pascal discovered it independently. In the western world we refer to it as Pascal's Triangle, but in China it is called Yang Hui's Triangle.

to p taken to the power of zero ($p^0 = 1$). The appropriate powers of q are seen to increase from left to right, going from $q^0 = 1$ up to q^3. Using that same procedure with the coefficients for $N = 4$ tosses, the binomial expansion looks like this:

$$(p + q)^4 = 1p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + 1q^4$$

When the appropriate values of p and q are inserted, this becomes the full mathematical expansion of the binomial. Suppose we are creating the binomial model for tossing a fair coin, with $p = ½$ and $q = ½$. The expansion of the binomial then returns the probability of each possible number of heads, from 0 to 4, that could be obtained in four tosses of the coin.

$$\begin{aligned}(p + q)^4 &= 1p^4 + 4p^3q + 6p^2q^2 + 6pq^3 + 1q^4 \\ &= \left(\frac{1}{2} + \frac{1}{2}\right)^4 = 1\left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) + 6\left(\frac{1}{2}\right)^2 q\left(\frac{1}{2}\right)^2 + 4\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 + 1\left(\frac{1}{2}\right)^4 \\ &= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16} = \frac{16}{16} = 1\end{aligned}$$

These outcomes of 0, 1, 2, 3, or 4 heads are mutually exclusive and exhaustive and their probabilities sum to one, so this binomial expansion constitutes a very simple probability model.

Figure 1.13 can be used to explain the construction of Pascal's triangle, that is, the triangle of the multiplicative coefficients that are used to create the binomial expansion at each power of $(p + q)$. The first row of the triangle has the single value "1." The next three rows are the coefficients of the first, second, and third powers of $p + q$ just calculated. To the left of the Pascal's triangle is shown a column labeled *N*. This indicates the power of $(p + q)$ for each row of the binomial coefficients. On the right side of the triangle is shown a column of the sums (*Sum*) of the coefficient values for each binomial power *N*. Notice that the sum for each row is the number 2 raised to the *N*th power. For example, the row of binomial coefficients corresponding to $N = 3$ (which reads "1 3 3 1") sums to 8 which is equal to 2 raised to the third power, 2^3. To create additional rows, the value of each coefficient is found by simply summing the two coefficients on the row above it: the one to the left above it plus the one to the right above it. For example, the third coefficient in the row corresponding to $N = 5$ is 10, which is the sum of the two coefficients 4 and 6 directly above it, with 4 on the left above and 6 on the right above. This particular Pascal's triangle has been constructed from an *N* of one toss up to an *N* of 10 tosses. Each of these rows of the triangle can be used to solve a complex probability problem as shown in what follows.

For an illustration of a simple applied probability model, suppose you are a student taking a psychometrics course. When you arrive at an early morning class you are confronted with a six-item multiple-choice pop quiz to test your knowledge of yesterday's readings. There are four alternatives (a, b, c, or d) to choose from on each of the six multiple-choice items. Unfortunately, you neglected to do the readings yesterday, so you will have to depend on the "luck of the draw" to pass the quiz. What is the probability of your getting a score of 66.7% or better using guesswork on this six-item quiz? Given that there are four alternatives on each item, the population probability p of getting a correct guess on any one of them is equal to ¼ under the null hypothesis that you do not know the correct answers. The probability q, of *not* getting the answer correct is $q = 1 - p = ¾$. Since there are $N = 6$ items, you select the row of binomial coefficients corresponding to $N = 6$ (which reads "1 6 15 20 15 6 1").

There are seven possible scores that you could obtain on this six-item quiz: 0, 1, 2, 3, 4, 5, or 6 items correct. To obtain the probability values for each of these possible scores, multiply each coefficient

by the population parameters of p and q raised to their appropriate power. Examining the "three-toss" binomial for $N = 3$ shown earlier, it was observed that the appropriate powers of p start at N and decrease by 1 for each term from left to right, from p^3 down to p taken to the power zero ($p^0 = 1$). The appropriate powers of q, on the other hand, are seen to increase from left to right, from q^0 to q^3. Applying that same procedure to the $N = 6$ coefficients, the binomial expansion algebraically looks like this:

$$(p + q)^6 = p^6 + 6p^5q + 15p^4q^2 + 20p^3q^3 + 15p^2q^4 + 6pq^5 + q^6$$

Entering the p and q values into the equation will return the probability values for each of the seven possible proportions (means of binary data) of correct answers. The highest possible proportion of correct answers is all six out of the six items, which corresponds to a proportion correct of 1.00. The lowest possible proportion is none of the six items answered correctly, which corresponds to a proportion correct of 0.00. Applying the algebraic formula to the seven named proportions correct, from "P(6 correct)" to "P(none correct)" creates the following calculated probabilities (P) for each of the seven proportions.

$$\begin{aligned}(p + q)^6 &= P(6\text{ correct}) + P(5\text{ correct}) + P(4\text{ correct}) + P(3\text{ correct}) + P(2\text{ correct}) \\ &\quad + P(1\text{ correct}) + P(\text{none}) \\ &= \left(\frac{1}{4}\right)^6 + 6\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right) + 15\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^2 + 20\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^3 + 15\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^4 + 6\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^5 + \left(\frac{3}{4}\right)^6 \\ &= {}^1\!/_{4096} + {}^{18}\!/_{4096} + {}^{135}\!/_{4096} + {}^{540}\!/_{4096} + {}^{1215}\!/_{4096} + {}^{1458}\!/_{4096} + {}^{729}\!/_{4096} \\ &= {}^{4096}\!/_{4096} = 1.00\end{aligned}$$

The values of the seven binary means (the proportions correct) are mutually exclusive values, they are exhaustive of all the possibilities for six-toss trials, and the probabilities of the seven outcomes sum to 1.00. These are the three properties that constitute a probability model.

Our question was "what is the probability of getting a score of 66.7% or higher on the pop quiz using guesswork?" Four out of the six answers correct is a score of ⅔ correct, that is 66.7% correct, so the probability of getting a score of 66.7% or higher is the sum of the three probabilities, P(4 correct), P(5 correct), and P(6 correct). These three probabilities sum to ${}^{154}\!/_{4096} = {}^1\!/_{4096} + {}^{18}\!/_{4096} + {}^{135}\!/_{4096}$, which is 0.0376, or a little less than 4 in 100. Not a very likely outcome.

Section §13, the final section, will now close with the introduction of the multinomial distribution and its use in demonstrating the logical proof and the mathematical basis of two foundational principles of inferential statistics that were introduced at the beginning of this section. The first is the principle expressed in Equation 1.4, that the variance of means is equal to one Nth of the variance of raw scores, where N is the number of observations averaged. This variance of the SDM can also be referred to as the *squared standard error of the mean*. The second is the formula for calculating an unbiased estimate of population variance from sample data by using degrees of freedom in the denominator as shown in Equation 1.2:

$$\hat{\sigma}^2 = \sum x_i^2/(N-1).$$

1.3.6.5 Multinomial Demonstrations of Two Fundamental Principles of Statistical Inference

At the beginning of this final sub-section of the chapter, stochastic modeling is taken to a higher level by introducing the *multinomial distribution probability model*. This is similar in structure to the binomial distribution, but whereas the binomial is appropriate only for binary data, the multinomial distribution may be used with *discrete* quantitative data. The first of the two fundamental

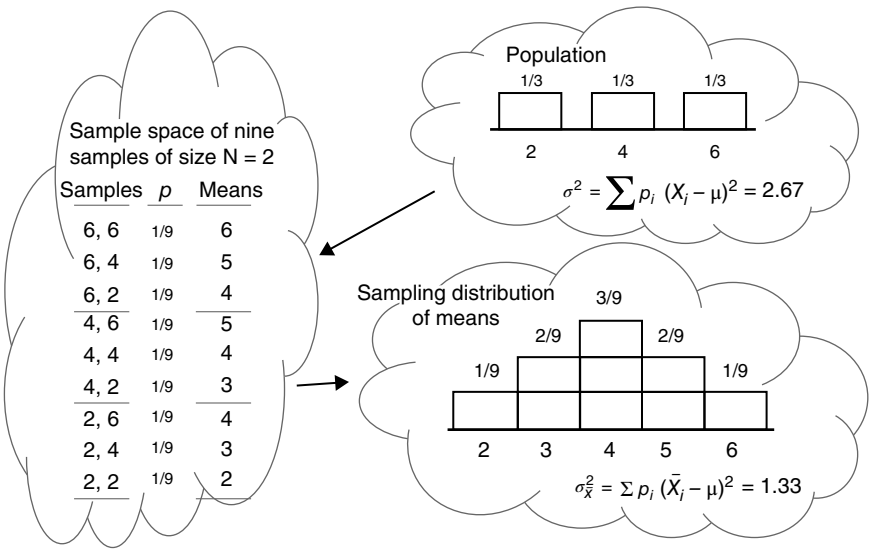


Figure 1.14 Diagram of the process of creating an SDM from a multinomial quantitative population distribution: (1) create a sample space of all the possible samples of size *N* from the population, with the probability and mean of each; (2) create a histogram displaying each of the means, labeled with its respective probability. (3) Calculate the variance of the SDM.

principles to be demonstrated is contained in Equation 1.4, the calculation of the variance of the SDM as one *N*th of the variance of raw scores, $\hat{\sigma}^2_{\overline{X}} = \hat{\sigma}^2/N$. This variance of the SDM, $\hat{\sigma}^2_{\overline{X}}$, is also the *squared standard error of the mean*, as stated in footnote 20 at the beginning of Section §13.

The first step in this demonstration is to create the simplest case multinomial population. This particular multinary population has only three values, 2, 4, and 6 as shown in Figure 1.14. Suppose these are pipes produced in a factory where there are equal numbers of pipes of each length produced. Each pipe length has a probability of $p = 1/3$ within the population. Since 2″, 4″, and 6″ pipes are equally likely in this population of pipes, the mean length is $\mu = 4$. The raw score variance is calculated using Equation 1.11. This is the "algebra of expectations" equation for a variance introduced in Appendix 2.

$$\sigma^2 = \sum p_i\,(X_i - \mu)^2 \qquad (\textit{Equation 1.11})$$

$$\sigma^2 = \sum p_i\,(X_i - \mu)^2 = \left(\frac{1}{3}\right)(2-4)^2 + \left(\frac{1}{3}\right)(4-4)^2 + \left(\frac{1}{3}\right)(6-4)^2 = \frac{4}{3} + \frac{0}{3} + \frac{4}{3} = \frac{8}{3} = 2.67$$

The next step of the demonstration is to create the sample space of all possible samples of a given size from this population, including the probability (p) for each sample.[31] For the sake of simplicity of exposition, a sample size of $N = 2$ is used to create a sample space of only nine unique sample configurations, as shown on the left side of Figure 1.14. The third column in this sample space listing is the mean of each of the six possible means. There are five possible mean values, 2, 3, 4, 5 and 6, and their respective probabilities are 1/9, 2/9, 3/9, 2/9, and 1/9 as shown in the lower right of Figure 1.14. This is because each of the nine samples has a probability of 1/9, and there are three samples that have a mean of 4, two samples that have a mean of 3, two samples that have a mean

31 Since each sample has two pipes, such as 'two 6″ pipes in the first sample', 'a 6″ pipe and a 4″ pipe in the second sample', etc., and since each pipe length has a probability of 1/3, using the multiplication rule of "AND" probabilities, Equation 1.9, the probability value for each of the nine samples is 1/9.

of 5, one sample that has a mean of 6, and one that has a mean of 2. This creates the pyramid shape that can be seen for the SDM for these nine samples shown in the lower right of Figure 1.14.

The population has a mean of $\mu = 4$, and this is, of course, also the mean of the SDM $\mu_{\overline{x}} = 4$. Equation 1.12 can now be used to calculate the variance of the SDM, $\sigma^2_{\overline{X}}$.

$$\sigma^2_{\overline{X}} = \sum p_i \left(\overline{X}_i - \mu\right)^2 = \left(\frac{1}{9}\right)(2-4)^2 + \left(\frac{2}{9}\right)(3-4)^2 + \left(\frac{3}{9}\right)(4-4)^2 + \left(\frac{2}{9}\right)(5-4)^2 + \left(\frac{1}{9}\right)(6-4)^2$$

$$= \left(\frac{4}{9}\right) + \left(\frac{2}{9}\right) + \left(\frac{0}{9}\right) + \left(\frac{2}{9}\right) + \left(\frac{4}{9}\right) = \frac{12}{9} = 1.33 \qquad \textit{(Equation 1.12)}$$

The principle being demonstrated here is that the variance of the SDM is equal to one *N*th of the variance of raw scores in the population, where *N* is sample size, that is, the number of values being averaged to create the means. The variance of raw scores in this population of pipe lengths is $\sigma^2 = 2.67$, and the sample size for creating the SDM is $N = 2$. The variance of the SDM for a sample size of two should therefore be half of the variance of raw scores in the population as it is shown to be here: the population variance of raw scores is two and two-thirds and the variance of the SDM is exactly half that, one and one third. This same demonstration could be done with a sample size of $N = 3$ or of $N = 4$. Even a sample size of three would be quite tedious, with 27 unique samples in the sample space. A sample size of four would have 81 unique samples in the sample space. Rest assured that were this to be done, the variance of the SDM for a sample size of $N = 3$ would be $\hat{\sigma}^2_{\overline{X}} = \hat{\sigma}^2/N = 2.667/3 = 0.888$, the variance of the SDM for a sample size of $N = 3$ would be $\hat{\sigma}^2_{\overline{X}} = \hat{\sigma}^2/N = 2.667/4 = 0.666$, and if it were to be done for a sample size of $N = 10$, it would be $\hat{\sigma}^2_{\overline{X}} = \hat{\sigma}^2/N = 2.667/10 = 0.267$.

The second demonstration of a foundational principle in statistical inference using the multinomial will now be given. In the paragraphs leading up to Table 1.3 at the beginning of Section §13 two formulas for a variance were given. The first was Equation 1.1, $S^2 = \sum x_i^2/N = \left[\Sigma\left(X_i - \overline{X}\right)^2\right]/N$, where the sum of squared deviations from the mean was divided by *N*, the sample size to obtain an ordinary descriptive sample variance. The second formula, given in Equation 1.2, is the correct formula to use for statistical inference. This formula will provide an unbiased point estimate of the true population variance. This formula, $\hat{\sigma}^2 = \sum x_i^2/(N-1) = \left[\Sigma\left(X_i - \overline{X}\right)^2\right]/(N-1)$, uses degrees of freedom, $N - 1$, in the denominator in calculating variance rather than *N*.

It was there stated that Equation 1.1 is a biased estimator and that if it were used as the point estimate of the population variance, it would systematically underestimate the true population variance σ^2 by "one *N*th" of its value. That statement is exactly correct, as will now be demonstrated in Figure 1.15 and the accompanying explanation.

Figure 1.15 uses the same population of pipe lengths that was introduced with Figure 1.14, which has a population mean of $\mu = 4$ and a population variance of $\sigma^2 = 2.67$. The listing of the nine samples of size two, $N = 2$, shown in the leftmost column of Figure 1.15 will immediately be recognized as the same nine samples that were shown in the Figure 1.14 demonstration. Since they are the same nine samples, they will also each have the same sample means, and each of these nine samples has the same probability of $p = 1/9$ as those. The second column in Figure 1.15 gives the sum of squared deviations for each of the samples, that is, the numerator that will be used in calculating the variances. The first sample, "6, 6" obviously has a mean of 6 and the two scores have zero deviations from their mean. Therefore the sum of squared deviations is 0. The same is true of the fifth sample,

Listing of the nine N = 2 samples in the sample space	Sum of squares for each sample	Sample variance (S^2) from each sample	Unbiased estimate of population variance ($\hat{\sigma}^2$) from each sample
6, 6	0	0	0
6, 4	2	1	2
6, 2	8	4	8
4, 6	2	1	2
4, 4	0	0	0
4, 2	2	1	2
2, 6	8	4	8
2, 4	2	1	2
2, 2	0	0	0

Expected value of the sample variances across the nine samples

$$E(S^2) = \sum p_i S_i^2 = (3/9)(0)+(4/9)(1)+(2/9)(4) = 4/9 + 8/9 = 12/9 = 1.33$$

Expected value of the unbiased estimates of population variance across the nine samples

$$E(\hat{\sigma}^2) = \sum p_i \hat{\sigma}_i^2 = (3/9)(0)+(4/9)(2)+(2/9)(8) = 8/9 + 16/9 = 24/9 = 2.67$$

Figure 1.15 Multinomial demonstration comparing the expected value of the sampling distribution of S^2 (descriptive sample variances) with the expected value of the sampling distribution of $\hat{\sigma}^2$ (unbiased estimates of population variance).

"4, 4," and the ninth sample "2, 2." All three samples have zero deviations from their mean and therefore a sum of squared deviations of 0. Since this zero sum of squares happens three times out of the nine samples, the probability connected with the sum of squares value of $SS_1 = 0$ is $p_1 = 3/9$.

Examining the second column, it can be seen that four samples have the sum of squares values of $SS = 2$. These are sample 2 ("6, 4") and sample 4 ("4, 6"), both of which have a mean of 5 and therefore deviation scores of 1 and −1, and therefore a sum of squares of these deviations of $SS_2 = 2$. The same is true of sample 6 ("4, 2") and sample 8 ("2, 4"), except that the two of them have means of 3, but they also have deviation scores of 1 and −1, and therefore a sum of squares of these deviations of $SS_2 = 2$. Since all four of these have the same sum of squares equal to 2, the probability connected with this sum of squares value of 2 is $p_2 = 4/9$.

Again examining the second column, it can be seen that only two samples have the sum of squares values of $SS = 8$. These are sample 3 ("6, 2") and sample 7 ("2, 6"), both of which have a mean of 4 and therefore deviation scores of 2 and −2, and therefore a sum of squares of these deviations of $SS_3 = 4 + 4 = 8$, and the probability connected with this sum of squares value of 8 is $p_3 = 2/9$.

To summarize there are only three different sums of squares values among these nine samples, $SS_1 = 0$, $SS_2 = 2$, and $SS_3 = 8$, and their corresponding three probabilities are $p_1 = 3/9$, $p_2 = 4/9$, and $p_3 = 2/9$. The three descriptive sample variances are found by dividing these three sum of squares values by their N. The first one with a sum of squares of zero also has a descriptive sample variance of zero, $S_1^2 = \sum x_i^2/N = SS_1/N = 0/2 = 0$, and it happens three out of nine times, so its probability is $p_1 = 3/9$. The second one with a sum of squares of 2, has a descriptive sample variance of 1, $S_2^2 = SS_2/N = 2/2 = 1$, and it happens four out of nine times, so its probability is $p_2 = 4/9$. The third one with a sum of squares of 8, has a descriptive sample variance of 4, $S_3^2 = SS_3/N = 8/2 = 4$, and it happens two out of nine times, so its probability is $p_3 = 2/9$. The expected value, that is the mean, of these three descriptive sample variance outcomes should underestimate the true population variance by one Nth, which is by ½, since the N in each of these samples is two. That means it will be half as big as the true population variance. Since the true population variance is 2.67 (two and two-thirds), it should come out to 1.33 (one and one-third). Using the "expected values" formula for a mean from Appendix 2, that is exactly what it comes out to be:

$$E\left(\hat{S}^2\right) = \sum p_i S_i^2 = \left(\frac{3}{9}\right)(0) + \left(\frac{4}{9}\right)(1) + \left(\frac{2}{9}\right)(4) = \frac{4}{9} + \frac{8}{9} = \frac{12}{9} = 1.33$$

The unbiased estimates of the true population variance are found by dividing these three sum of squares values by $N-1$, their degrees of freedom. The first one with a sum of squares of zero also has an unbiased variance estimate of zero, $S_1^2 = \sum x_i^2/(N-1) = SS_1/(N-1) = 0/1 = 0,$ and it happens three out of nine times, so its probability is $p_1 = 3/9$. The second one with a sum of squares of 2, has an unbiased variance estimate of 2, $S_2^2 = SS_2/(N-1) = 2/1 = 2,$ and it happens four out of nine times, so its probability is $p_2 = 4/9$. The third one with a sum of squares of 8, has an unbiased variance estimate of 4, $S_3^2 = SS_3/N = 8/2 = 4,$ and it happens two out of nine times, so its probability is $p_3 = 2/9$. The mean, that is the expected value of these three unbiased variance estimate outcomes should agree with the true population variance of 2.67, which is exactly what it comes out to be:

$$E(\hat{\sigma}^2) = \sum p_i \hat{\sigma}_i^2 = \left(\frac{3}{9}\right)(0) + \left(\frac{4}{9}\right)(2) + \left(\frac{2}{9}\right)(8) = \frac{8}{9} + \frac{16}{9} = \frac{24}{9} = 2.67$$

Q. E. D., *quod erat demonstrandum*.

1.4 Summary

Chapter 1 begins with a discussion of why mathematical methods are central to the empirical sciences and particularly why statistical methods are used in the behavioral sciences and why we would use statistics rather than just a branch of applied mathematics. It then discusses the historical relationship between psychology and the emergence of statistical methods and how the Pearson product-moment correlation coefficient emerged from the work of Galton (perhaps the first psychometrician) and Pearson on the heritability of intelligence.

Section §6 then gives an account of the ongoing debate of the Ferguson Committee of the British Association for the Advancement of Science regarding whether the social and behavioral sciences could be considered to have bona fide scientific measurement, and a series of answers to that challenge. The first answer to the challenge came in the form of the influential 1946 paper by Stevens proposing four kinds of measurement scales, followed by a strong rebuttal to the Ferguson Committee challenge from Luce and Tukey (1964) using ACM, which in turn led to a series of three volumes on the foundations of measurement by Krantz, Luce, Suppes, and Tversky (1971, 1989, 1990). A brief account is also given of some encouraging and positive recent developments in the more general discipline of metrology and its most recent approach to measurement across the sciences.

The last part of the chapter deals with the development of scientific variables and the place of stochastic methods, comparisons of measures of central tendency and of dispersion in the development of psychometric methods. The chapter ends with a review of principles of probability and statistical inference that are foundational for psychometric work.

1.5 Study Questions

A. Discussion Questions

Q1 What is the foundational quantitative method for the past 120 years of development of psychometric theory and practice? Who was the author? *(the third paragraph of the introduction to Chapter 1)*

Q2 What multivariate analytical method was invented four years earlier in 1901 that led directly to the method described in question #1? Who was the author *(the third paragraph of the introduction to Chapter 1)*

§1 Why Are Mathematical Methods Central to the Empirical Sciences? Why Are Statistical Methods Used in the Behavioral Sciences? (p. 2)

Q3 Why have statistical methods become the primary quantitative approach within the behavioral sciences? *(all of Section **§1**)*

§2 The Specific Mission of Statistical Methods in the Behavioral Sciences (p. 3)

Q4 Explain the fundamental scientific functions of explanation and prediction. How do they differ and how does each contribute? *(Section **§2**, paragraphs 2 through 4)*

Q5 Compare the twin roles of statistical methods in scientific research, the exploratory role and the confirmatory role. How does science benefit from each? *(Section **§2**, paragraphs 2 and 3)*

§3 Why Do We Use *Behavioral Statistics* and Not Just a Branch of Applied Mathematics? (p. 4)

Q6 Explain the "historical development and mutual influence" of psychology and statistical science. *(all of Section **§3**)*

§4 Some Remarks on the Historical Relationships Between Psychology and Statistics: The Development of the Correlation Coefficient and Least Squares Regression (p. 5)

Q7 Who were Mendel and Galton? What is the difference between the Galtonian and the Mendelian mathematical approaches to genetics, as discussed by Sir R. A. Fisher (1918)? Describe how each of these approaches was carried out. *(Section **§4**, first paragraph)*

Q8 How was the *Pearson product-moment correlation coefficient* (PPMCC) developed? Who was its author? What were its historical antecedents? In what year was it published? *(Subsection **1.1.4.1**, paragraphs 1–2)*

Q9 When was the "least squares" mathematical method discovered, and by whom? Which very widely used contemporary statistical method grew out of it in the 20th and 21st centuries? *(Subsection **1.1.4.1**, third paragraph)*

Q10 Why is the symbol "*r*" used to represent the PPMCC? What does it stand for? *(Subsection **1.1.4.2**, fifth paragraph)*

Q11 To whom did Pearson initially attribute the conceptual/mathematical work that eventually led to his correlation coefficient? To whom did he finally give the credit in his 1920 *Biometrika* paper? Why did he change his mind? *(Subsection **1.1.4.2**, eighth paragraph)*

§5 On Terminology I: Attributes with Categories—Variables with Values—Tests with Scores. And What About "Data"? (p. 10)

Q12 Comment on each of the following three areas of semantic ambiguity identified by the authors in the terminology of behavioral research methods: the concept of "score," the

concept of "variable," and the concept of "data." *(first paragraph of Section **§5**; all of Subsection **1.2.1.1**, especially paragraphs 1, 4, 6, and 8)*

Q13 Summarize the three broad preconditions identified by the authors for producing bona fide research data. *(all of Subsection **1.2.1.2**)*

§6 Measurement Scales and the Representational Theory of Measurement (p. 13)

Q14 Explain how S. Smith Stevens' (1946) landmark paper on the four types of measurement, that is, the four measurement scales, grew out of the seven-year debate of the Ferguson Committee of the British Association for the Advancement of Science. *(all of Subsection **§6a** and Subsection **1.2.2.2**)*

Q15 What is meant by the term "ostensive fundamental measurement"? *(footnote 7 in Subsection **1.2.2.1**)*

Q16 Give data examples that fit each of Stevens' four types of measurement and identify the properties that distinguish among the four types. *(all of Subsection **1.2.2.3**, particularly the first five paragraphs and Figure 1.1)*

Q17 At which of the four measurement scales does Stevens say that "we come to a scale that is quantitative in the ordinary sense of the word?" On this basis, at what point should one place a "major dividing line" through the four scale types? *(Subsection **1.2.2.3**, the fourth paragraph)*

Q18 What is the central idea of the "relational theory of measurement" (RTM)? How did it evolve from its early beginnings in the late 19th century to the relational theory of Norman Campbell, to Stevens' popular redefinition as a theory of scale types, to its highly technical development in the axiomatic relational theory of measurement? *(Subsection **1.2.2.4**, particularly the first four paragraphs)*

Q19 What do Scott and Suppes (1958) and Domotor (1992) mean by "imbedding" and "injection"? What do they mean by projection? *(Subsection **1.2.2.4**, paragraphs 5 through 9).*

Q20 Compare the axiomatic RTM approach measurement to the huge academic industry of mainstream psychometric theory and practice including factor analysis, "true score theory," and CTT. How complex and conceptually challenging is each? How popular and practically useful is each? Where has each had its major impact? *(Subsection **1.2.2.5**)*

Q21 What is metrology and how has it affected the debate concerning the necessity of concatenation operations being essential to true scientific measurement? *(Subsections **1.2.2.6** and **1.2.2.4**)*

§7 *Scientific* and *Stochastic* Variables: When May "Data" Be Treated By Statistical and Test Theory Methods? (p. 29)

Q22 In what way is the analysis of test scores into the two components of "true score" and "measurement error," the stochastic component, central to CTT? *(Section **§7**, first paragraph)*

Q23 What is a random variable? What is the value of a truly random variable in quantifying the stochastic component of data? *(Section **§7**, paragraphs 2 and 3)*

Q24 In what way can a random process such as tossing two dice create a precisely specified distribution for the error component? *(Section **§7**, paragraphs 2 and 3, and Figure 1.6)*

Q25 Explain the *Central Limit Theorem* and how it relates to the sampling distribution of the "sums of the two dice values" shown on page 22. *(Section **§7**, paragraph 3, and Figure 1.6)*

Q26 What are the three necessary conditions for creating a systematic and precise probability model? *(Section **§7**, paragraphs 4 and 5, and Figure 1.6)*

§8 Group Level of an Attribute: The Location of the Group's Scores on an Interval Scale (p. 31)

Q27 How do histograms and means help in the description and comparison of two distributions? *(Section **§8**, paragraphs 1 to 3)*

Q28 Describe the two ways of calculating a mean, the traditional way of summing the scores and dividing by N, and the probabilistic way of multiplying each score by its "relative frequency," i.e. its probability of summing the products. *(Section **§8**, paragraphs 3 and 4)*

Q29 Describe the two ways of interpreting the fathers' mean head circumference (X) and the sons' mean head circumference (Y): the empirical/descriptive, and the formal/mathematical. *(Section **§8**, paragraphs 5 through 7)*

§9 Comparing median versus mean: Which one is "fair"? (p. 34)

Q30 What are the relative advantages and disadvantages of the mean and the median as measures of central tendency? *(Section **§9**, paragraphs 1 and 2)*

§10 A Psychometric Look at the Mean $\overline{X}$: Item *Difficulty* and p (p. 35)

Q31 Describe how the "oversimplified explanation of a performance score" is based on two concepts: the _____ of the tested subject, and the _____ of the task. How are these put together to create a dominance relation and to make inferences regarding the subject and the task? *(Section **§10**, paragraphs 1 and 2)*

Q32 How can a psychometric testing study be interpreted in two ways, as an agglomerative *attribute of a group of people*, or as *a property of the method, that is, of the test itself*? *(Section **§10**, paragraphs 2 through 6)*

Q33 How is a proportion calculated on a "dummy variable" (binary) related to the mean of a fully quantitative variable? *(Section **§10**, paragraphs 7 through 10)*

§11 Intra-group and Inter-individual Variability—*Dispersion* of Scores (p. 37)

Q34 Why are the standard deviation and variance the primary statistics for quantifying dispersion rather than the range? How are the range's characteristics like those of the median? *(Section **§11**, paragraph 4)*

Q35 Why are the standard deviation and variance the primary statistics for quantifying dispersion rather than MAD? *(Section **§11**, paragraphs 4 to 6)*

Q36 Why is the squaring of values a better way to deal with the problem of negative values rather than using absolute values? *(Section **§11**, paragraphs 6 and 7)*

§12 A Small Numerical Example, Finally (p. 40)

Q37 What is the approximate percent of the area under the normal curve found between plus and minus one standard deviation units from the mean? What is the approximate percent of the area between two and minus two standard deviation units? *(Section **§12**, paragraphs 3 and 4)*

§13 Does the Larger Head Circumference in a Sample of Sons than Fathers Hold for the Whole Population? A Preliminary Sidestep to Statistical Inference? (p. 40)

Q38 How is the unbiased point estimate of the population standard deviation calculated? How is it useful? Why is it so important to get the calculation right? *(Section **§13**, paragraphs 5 through 8, and Equation 1.3)*

Q39 How is the standard error of the mean calculated? How is it useful? *(Subsection **1.3.6.1**, paragraphs 1 through 4,and Equations 1.5 and 1.5a)*

Q40 What is a confidence interval for the point estimate of the population mean and how is it calculated? How is it useful? *(Subsection **1.3.6.2**, paragraphs 5 to 7, and Equation 1.6)*

Q41 How is the standard error of the proportion calculated? What is unusual about its calculation when compared to the standard error of the mean? *(Subsection **1.3.6.2**, paragraphs 1 through 5, Equation 1.8, and footnote 23)*

Q42 Explain the multiplication rule for independent joint events. How is it used in creating the binomial distribution? *(Subsection **1.3.6.2**, paragraphs 6 through 9, Equation 1.9 shown for both two and three joint events, Figures 1.9 and 1.10)*

Q43 Explain the addition rule for mutually exclusive composite events. How is it used in creating the binomial distribution? *(Subsection **1.3.6.2**, paragraphs 10 through 19, Equation 1.10 for two composite events and Equation 1.10a for three composite events, footnotes 27 and 28)*

Q44 How does the binomial distribution help to clarify the rationale and logic of an SDM? *(Subsection **1.3.6.3**, paragraphs 1 through 7, Figure 1.12)*

Q45 How is the SDM useful in statistical inference? *(Subsection **1.3.6.4**, paragraphs 1 through 8, Figure 1.13 and Figure 1.14)*

B. Exercise Questions

EQ1 A psychometrics class teacher gives a short eight-item multiple choice quiz at the beginning of class each morning to ensure students have read the material. The student must get a "pass" for each day by getting six or more of the eight items correct (75%) when choosing from among three alternative answers on each question. The binomial distribution can determine the probability of a student receiving a pass by random guessing alone. (The resources for building the binomial distribution for answering this question are found in Figure 1.3 and the accompanying text. Since there are 8 items in the quiz, go to the column labeled ***N***, then go to the row with an N of 8. Copy the nine coefficients to the right of it. The first of the nine is the coefficient for "0" items correct, the second for '1' items correct, the third for '2' items correct,...and the ninth for 'all 8 items correct.' These nine coefficients account for all nine of the possible outcomes in tossing a coin eight times or in our case guessing the answer to six or more items by chance. To construct the binomial distribution probabilities for each of the nine possible outcomes it is necessary to multiply *the* coefficient for each possible number of items 'guessed' correctly by its power of p and its power of q for each coefficient as outlined in Subsection ***1.3.6.3***, using the multiplication rule and the addition rule.)

EQ2 Use a spreadsheet to create a *Pascal's Triangle Calculator* following the instructions below. Use it to create the "bar graph of the sampling distribution of the proportion of heads for twenty-toss trials of a fair coin" ($p = 1/2$ and $q = 1/2$) as shown in the upper right panel of Figure 1.11 and compare your results with those in the figure."

Instructions for creating a *Pascal's Triangle Calculator* on a spreadsheet. This can be used for calculating binomial coefficients for binomial distributions with N values up to 20.

Use a spreadsheet such as Excel to create a Pascal's triangle calculator following the instructions below.

1. *Place a 1 in each cell of column A of a spreadsheet such as Excel, from cell A1 to cell A21.*
2. *Create a diagonal line of ones by going "down one and over right one" repeatedly from cell B2 to cell U21. The progression goes like this: B2, C3, D4, E5, ... down to ... Q17, R18, S19, T20, and U21. These two "lines" together with an underline across the bottom of row 21 will form a triangle that you can fill in.*
3. *Now begin to create the Pascal's triangle coefficients, beginning by entering the following into cell B3: "=A2 + B2". After hitting the "return" key, the number "2" should appear.*
4. *Copy the cell B3 you have just entered and paste its contents into row 4; cell B4 and then cell C4. A "3" should appear in each of them.*
5. *Now do the same thing to row 5, from B5 to D5. The numbers "4 6 4" should appear.*
6. *Continue doing this until you have completely filled the triangle.*

Now that you have successfully created your Pascal's Triangle Calculator, try it out by recreating the bar graph shown in the upper right corner of Figure 1.11 with a sample size of N = 20 and also with p = ½ and q = ½, a fair coin.

EQ3 *Now that you have created your Pascal's triangle calculator in Excel, and successfully used it to re-create the bar graph of the sampling distribution of the proportion of heads for 20-toss trials*

of a fair coin, shown in the upper right corner of Figure 1.11, do the same thing again, but this time with a weighted coin ($p = 1/4$ and $q = 3/4$) and compare to the lower right panel of Figure1.11. How does this sampling distribution of coin tossing with a p value of $p = 1/4$ compare to the one in the lower right corner of Figure 1.11 with a p value of $p = 1/3$?". Please use the same six-step instructions used for question EQ2 to now create this new Pascal's triangle for EQ3.

References

Alder, K. (2021). *The measure of all things: The seven-year odyssey and hidden error that transformed the world*. New York: Free Press.

Allen, M. J., & Yen, W. Y. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

Bravais, A. (1846). Analyse mathématique sur les probabilités des erreurs de situation d'un point. *Mémoires présentés par divers savants à l'Académie Royale des Sciences de l'Institut de France*, 9, 255–332.

BIPM (2019). (Bureau International des Poids et Mesures). *The international system of units (SI), 9th Edition*. St. Cloud, Paris, France: The International Bureau of Weights and Measures (BIPM). This SI Brochure is distributed under the terms of the Creative Commons Attribution 3.0 IGO https://creativecommons.org/licenses/by/3.0/igo/.

Brown, B. L., Hendrix, S. B., Hedges, D. W., & Smith, T. B. (2012). *Multivariate analysis for the biobehavioral and social sciences: A graphical approach*. Hoboken: John Wiley & Sons.

Brown, R. J. C. (2021). Measuring measurement – what is metrology and why does it matter? *Measurement (London)*, 168: 108408. https://doi.org/10.1016/j.measurement.2020.108408.

Campbell, N. R. (1920). *Physics: The elements*. Cambridge: Cambridge University Press. Reprinted (1957) as *Foundations of science: The Philosophy and theory of experiment*. New York: Dover.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London: Longmans.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (2021). *The twenty-first mental measurements yearbook*. Lincoln: University of Nebraska Press.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, 3(3), 186–190.

Coombs, C. H. (1964). *A theory of data*. New York: John Wiley & Sons, Inc.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Daston, L. (1995). *Classical probability in the enlightenment*. Princeton, NJ: Princeton University Press.

Domotor, Z. (1992). Measurement from empiricist and realist points of view. In Savage, C. W. & Ehrlich, P. *Philosophical and foundational issues in measurement theory*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Ferreiro, L. D. (2011). *Measure of the Earth: The enlightenment expedition that reshaped our world*. New York: Basic Books.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52, 399–433.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. London: Macmillan & Co.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society*, 45, 135–145.

Galton, F. (1889). *Natural inheritance*. London: Macmillan.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons, Inc.

Harman, H. H. (1960). *Modern factor analysis*. Chicago: The University of Chicago Press.

Helmholtz, H. (1887/1971). An epistemological analysis of counting and measurement. In Kahl, R. (Ed.), *Selected writings of Hermann von Helmholtz.* Connecticut: Wesleyan University Press.

Hilbert, D. (1899/2015). *Grundlagen der geometrie*. Berlin: Springer Spektrum.

Hölder, O. (1901). Die axiome der quantitiit und die lehre vom mass. *Berichte uber die verhandlungen der Koniglichen Sachsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, 53, l–64.

Horst, P. (1965). *Factor analysis of data matrices*. New York: Holt-Rinehart & Winston.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, 2, 389–423.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2), 263–291

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Volume I: Additive and polynomial representations*. New York: Academic Press.

LaPlace, P. S. (1814/1952). A philosophical essay on probabilities. In F. W. Truscott, & F. L. Emory (Ed.), *Translated from the 6th French*. New York: Dover Publications.

Laplace, P. S. (1847/1995). *Théorie analytique des probabilités*. Paris: Editions J. Gabay, pp. 3–4.

Legendre, A. M. (1805). *Nouvelles methodes pour la determination des orbites des cometes*. Paris: Courcier.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Boston: Addison-Wesley.

Luce, R. D., Krantz, D. M., Suppes, P., & Tversky, A. (1990). *Foundations of measurement Volume III: Representation, axiomatization, and invariance*. New York: Academic Press.

Luce, R. D., & Narens, L. (1994). Fifteen problems concerning the representational theory of measurement. *Patrick Suppes: Scientific Philosopher*, 2, 219–249.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.

Majerova, V., & Majer, E. (1999). *Kvalitativni vyzkum v sociologii venkova a zemedelstvi. Cast I*. Praha, Czech: Ceska zemedelska univerzita. (Qualitative research in sociology of rural areas and agriculture. Part I.).

Michell, J. (1990). *An introduction to the logic of psychological measurement*. London: Routledge.

Michell, J. (2020). Thorndike's *Credo*: metaphysics in psychometrics. *Theory and Psychology*, 30(3), 309–328.

Mislevy, R. J., Almond, R. G., & Lucas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Report 632). Los Angeles, CA: CRESST.

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). Menlo Park, CA: SRI International.

Narens, L. (2002). *Theories of meaningfulness*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2010). The Rasch model and additive cojoint measurement. In M. L. Garner, G. Engelhard, Jr., W. P. Fisher, Jr., M. Wilson (Eds.), *Advances in Rasch measurement, Volume 1*. Maple Grove, MN: JAM Press.

Newell, D, & Tiesinga, E. (2019). *The international system of units (SI), 2019 edition*. US Department of Commerce, National Institute of Standards and Technology.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1), 1–18.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society*, 187-A, 253–318.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1), 25–45.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6* 2(11), 559–572.

Peirce, C. S. (1901/1957). The logic of abduction. In V. Tomas (Ed.), *Essays in the philosophy of science*. Indiana, IN: Bobbs-Merill.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3, 237–255.

Pfanzagl, J. (1959). *Die axiomatischen grundlagen einer allgemeinen theorie des messens. Schriftenreihe Statist. Inst. Univ. Wien, Vol. 1*. Würzburg: Physica-Verlag.

Pfanzagl, J. (1968). *Theory of measurement*. New York: Wiley.

Piovani, J. I. (2008). The historical construction of correlation as a conceptual and operative instrument for empirical research. *Quality & Quantity*, 42, 757–777.

Plackett, R. L. (1972). Studies in the history of probability and statistics. XXIX: The discovery of the method of least squares. *Biometrika*, 59(2), 239–251.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Savage, C. W., & Ehrlich, P. (1992). *Philosophical and foundational issues in measurement theory*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Scott, D., & Suppes, P. (1958). Foundational aspects of theories of measurement. *The Journal of Symbolic Logic*, 23(2), 113–128.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.

Spearman, C. (1904b). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15, 201–292.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 18, 161–169.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, 3, 271–395.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology*, 5, 417–426.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680.

Stevens, S. S., & Davis, H. (1938). *Hearing: Its psychology and physiology*. New York: Wiley.

Stewart, I. (2012). *In pursuit of the unknown: 17 equations that changed the world*. New York: Basic Books.

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 9, 465–474.

Stock, M., Davis, R., de Mirandés, E., & Milton, M. J. (2019). The revision of the SI—the result of three decades of progress in metrology. *Metrologia*, 56, 022001.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, E. Galanter (Eds.), *Handbook of mathematical psychology, Volume I*. New York: Wiley.

Suppes P., Krantz, D. H., Luce R. D., & Tversky, A (1989). *Foundations of measurement, Volume II; Geometrical, threshold, and probabilistic representations*. New York: Academic Press.

Thurstone, L. L. (1931). *The reliability and validity of tests: Derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems*. Ann Arbor: Edwards Brothers, Inc. (This is now unavailable.)

Thurstone, L. L. (1935). *The vectors of mind*. Chicago: The University of Chicago Press.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: The University of Chicago Press.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement*, 16, 8–13.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. *Science*, 185, 1124–1131.