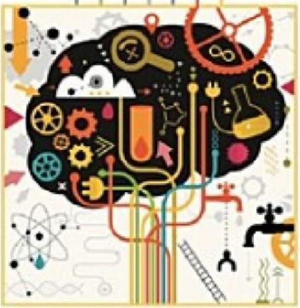
The Seven Pillars of Statistical Wisdom



STEPHEN M. STIGLER

Jacket illustration: DigitalVision Vectors, DrAfter123, royalty free/Getty Images
Jacket design: Jill Breitbarth

978-0-674-08891-7 (pbk. : alk. paper)
978-0-674-97021-2 (EPUB)
978-0-674-97020-9 (MOBI)

The Library of Congress has cataloged the printed edition as follows:
Names: Stigler, Stephen M., author.
Title: The seven pillars of statistical wisdom / Stephen M. Stigler.
Description: Cambridge, Massachusetts : Harvard University Press, 2016. | Includes bibliographical references and index.
Identifiers: LCCN 2015033367
Subjects: LCSH: Statistics—History. | Mathematical statistics—History.
Classification: LCC QA276.15 S754 2016 | DDC 519.5—dc23
LC record available at http://lccn.loc.gov/2015033367

CONTENTS

Introduction

1 **AGGREGATION** From Tables and Means to Least Squares

2 **INFORMATION** Its Measurement and Rate of Change

3 **LIKELIHOOD** Calibration on a Probability Scale

4 **INTERCOMPARISON** Within-Sample Variation as a Standard

5 **REGRESSION** Multivariate Analysis, Bayesian Inference, and Causal Inference

6 **DESIGN** Experimental Planning and the Role of Randomization

7 **RESIDUAL** Scientific Logic, Model Comparison, and Diagnostic Display

Conclusion

Notes
References
Acknowledgments
Index

· INTRODUCTION ·

What is Statistics? This question was asked as early as 1838—in reference to the Royal Statistical Society—and it has been asked many times since. The persistence of the question and the variety of answers that have been given over the years are themselves remarkable phenomena. Viewed together, they suggest that the persistent puzzle is due to Statistics not being only a single subject. Statistics has changed dramatically from its earliest days to the present, shifting from a profession that claimed such extreme objectivity that statisticians would only gather data—not analyze them—to a profession that seeks partnership with scientists in all stages of investigation, from planning to analysis. Also, Statistics presents different faces to different sciences: In some applications, we accept the scientific model as derived from mathematical theory; in some, we construct a model that can then take on a status as firm as any Newtonian construction. In some, we are active planners and passive analysts; in others, just the reverse. With so many faces, and the consequent challenges of balance to avoid missteps, it is no wonder that the question, "What is Statistics?" has arisen again and again, whenever a new challenge arrives, be it the economic statistics of the 1830s, the biological questions of the 1930s, or the vaguely defined "big data" questions of the present age.

With all the variety of statistical questions, approaches, and interpretations, is there then no core science of Statistics? If we are fundamentally dedicated to working in so many different sciences, from public policy to validating the discovery of the Higgs boson, and we are sometimes seen as mere service personnel, can we really be seen in any reasonable sense as a unified discipline, even as a science of our own? This is the question I wish to address in this book. I will not try to tell you what Statistics is or is not; I will attempt to formulate seven principles, seven pillars that have supported our field in different ways in the past and promise to do so into the indefinite future. I will try to convince you that each of these was revolutionary when introduced, and each remains a deep and important conceptual advance.

My title is an echo of a 1926 memoir, *Seven Pillars of Wisdom*, by T. E. Lawrence, Lawrence of Arabia.[1] Its relevance comes from Lawrence's own source, the Old Testament's Book of Proverbs 9:1, which reads, "Wisdom hath built her house, she hath hewn out her seven pillars." According to Proverbs, Wisdom's house was constructed to welcome those seeking understanding; my version will have an additional goal: to articulate the central intellectual core of statistical reasoning.

In calling these seven principles the Seven Pillars of Statistical Wisdom, I hasten to emphasize that these are seven *support* pillars—the disciplinary foundation, not the whole edifice, of Statistics. All seven have ancient origins, and the modern discipline has constructed its many-faceted science upon this structure with great ingenuity and with a constant supply of exciting new ideas of splendid promise. But without taking away from that modern work, I hope to articulate a unity at the core of Statistics both across time and between areas of application.

The first pillar I will call Aggregation, although it could just as well be given the nineteenth-century name, "The Combination of Observations," or even reduced to the simplest example, taking a mean. Those simple names are misleading, in that I refer to an idea that is now old but was truly revolutionary in an earlier day—and it still is so today, whenever it reaches into a new area of application. How is it revolutionary? By stipulating that, given a number of observations, you can actually gain information by throwing information away! In taking a simple arithmetic mean, we discard the individuality of the measures, subsuming them to one summary. It may come naturally now in repeated measurements of, say, a star position in astronomy, but in the seventeenth century it might have required ignoring the knowledge that the French observation was made by an observer prone to drink and the Russian observation was made by use of an old instrument, but the English observation was by a good friend who had never let you down. The details of the individual observations had to be, in effect, erased to reveal a better indication than any single observation could on its own.

The earliest clearly documented use of an arithmetic mean was in 1635; other forms of statistical summary have a much longer history, back to Mesopotamia and nearly to the dawn of writing. Of course, the recent important instances of this first pillar are more complicated. The method of least squares and its cousins and descendants are all averages; they are weighted aggregates of data that submerge the identity of individuals, except for designated covariates. And devices like kernel estimates of densities and various modern smoothers are averages, too.

The second pillar is Information, more specifically Information Measurement, and it also has a long and interesting intellectual history. The question of when we have enough evidence to be convinced a medical treatment works goes back to the Greeks. The mathematical study of the rate of information accumulation is much more recent. In the early eighteenth century it was discovered that in many situations the amount of information in a set of data was only proportional to the square root of the number *n* of observations, not the number *n* itself. This, too, was revolutionary: imagine trying to convince an astronomer that if he wished to double the accuracy of an investigation, he needed to quadruple the number of observations, or that the second 20 observations were not nearly so informative as the first 20, despite the fact that all were equally accurate? This has come to be called the root-*n* rule; it required some strong assumptions, and it required modification in many complicated situations. In any event, the idea that information in data could be measured, that accuracy was related to the amount of data in a way that could be precisely articulated in some situations, was clearly established by 1900.

By the name I give to the third pillar, Likelihood, I mean the calibration of inferences with the use of probability. The simplest form for this is in significance testing and the common *P*-value, but as the name "Likelihood" hints, there is a wealth of associated methods, many related to parametric families or to Fisherian or Bayesian inference. Testing in one form or another goes back a thousand years or more, but some of the earliest tests to use probability were in the early eighteenth century. There were many examples in the 1700s and 1800s, but systematic treatment only came with the twentieth-century work of Ronald A. Fisher and of Jerzy Neyman and Egon S. Pearson, when a full theory of likelihood began serious development. The use of probability to calibrate inference may be most familiar in testing, but it occurs everywhere a number is attached to an inference, be it a confidence interval or a Bayesian posterior probability. Indeed, Thomas Bayes's theorem was published 250 years ago for exactly that purpose.

The name I give the fourth pillar, Intercomparison, is borrowed from an old paper by Francis Galton. It represents what was also once a radical idea and is now commonplace: that statistical comparisons do not need to be made with respect to an exterior standard but can often be made in terms interior to the data themselves. The most commonly encountered examples of intercomparisons are Student's *t*-tests and the tests of the analysis of variance. In complex designs, the partitioning of variation can be an intricate operation and allow blocking, split plots, and hierarchical designs to be evaluated based entirely upon the data at hand. The idea is quite radical, and the ability to ignore exterior scientific standards in doing a “valid” test can lead to abuse in the wrong hands, as with most powerful tools. The bootstrap can be thought of as a modern version of intercomparison, but with weaker assumptions.

I call the fifth pillar Regression, after Galton’s revelation of 1885, explained in terms of the bivariate normal distribution. Galton arrived at this by attempting to devise a mathematical framework for Charles Darwin’s theory of natural selection, overcoming what appeared to Galton to be an intrinsic contradiction in the theory: selection required increasing diversity, in contradiction to the appearance of the population stability needed for the definition of species.

The phenomenon of regression can be explained briefly: if you have two measures that are not perfectly correlated and you select on one as extreme from its mean, the other is expected to (in standard deviation units) be less extreme. Tall parents on average produce somewhat shorter children than themselves; tall children on average have somewhat shorter parents than themselves. But much more than a simple paradox is involved: the really novel idea was that the question gave radically different answers depending upon the way it was posed. The work in fact introduced modern multivariate analysis and the tools needed for any theory of inference. Before this apparatus of conditional distributions was introduced, a truly general Bayes’s theorem was not feasible. And so this pillar is central to Bayesian, as well as causal, inference.

The sixth pillar is Design, as in “Design of Experiments,” but conceived of more broadly, as an ideal that can discipline our thinking in even observational settings. Some elements of design are extremely old. The Old Testament and early Arabic medicine provide examples. Starting in the late nineteenth century, a new understanding of the topic appeared, as Charles S. Peirce and then Fisher discovered the extraordinary role randomization could play in inference. Recognizing the gains to be had from a combinatorial approach with rigorous randomization, Fisher took the subject to new levels by introducing radical changes in experimentation that contradicted centuries of experimental philosophy and practice. In multifactor field trials, Fisher's designs not only allowed the separation of effects and the estimation of interactions; the very act of randomization made possible valid inferences that did not lean on an assumption of normality or an assumption of homogeneity of material.

I call the seventh and final pillar Residual. You might suspect this is an evasion, "residual" meaning "everything else." But I have a more specific idea in mind. The notion of residual phenomena was common in books on logic from the 1830s on. As one author put it, "Complicated phenomena … may be simplified by subducting the effect of known causes, … leaving … a *residual phenomenon* to be explained. It is by this process … that science … is chiefly promoted."[2] The idea, then, is classical in outline, but the use in Statistics took on a new form that radically enhances and disciplines the method by incorporating structured families of models and employing the probability calculus and statistical logic to decide among them. The most common appearances in Statistics are our model diagnostics (plotting residuals), but more important is the way we explore high-dimensional spaces by fitting and comparing nested models. Every test for significance of a regression coefficient is an example, as is every exploration of a time series.

At serious risk of oversimplification, I could summarize and rephrase these seven pillars as representing the usefulness of seven basic statistical ideas:

1. The value of targeted reduction or compression of data
2. The diminishing value of an increased amount of data
3. How to put a probability measuring stick to what we do

4. How to use internal variation in the data to help in that
5. How asking questions from different perspectives can lead to revealingly different answers
6. The essential role of the planning of observations
7. How all these ideas can be used in exploring and comparing competing explanations in science

But these plain-vanilla restatements do not convey how revolutionary the ideas have been when first encountered, both in the past and in the present. In all cases they have pushed aside or overturned firmly held mathematical or scientific beliefs, from discarding the individuality of data values, to downweighting new and equally valuable data, to overcoming objections to any use of probability to measure uncertainty outside of games of chance. And how can the variability interior to our data measure the uncertainty about the world that produced it? Galton's multivariate analysis revealed to scientists that their reliance upon rules of proportionality dating from Euclid did not apply to a scientific world in which there was variation in the data—overthrowing three thousand years of mathematical tradition. Fisher's designs were in direct contradiction to what experimental scientists and logicians had believed for centuries; his methods for comparing models were absolutely new to experimental science and required a change of generations for their acceptance.

As evidence of how revolutionary and influential these ideas all were, just consider the strong push-back they continue to attract, which often attacks the very aspects I have been listing as valued features. I refer to:

Complaints about the neglect of individuals, treating people as mere statistics

Implied claims that big data can answer questions on the basis of size alone

Denunciations of significance tests as neglectful of the science in question

Criticisms of regression analyses as neglecting important aspects of the problem

These questions are problematic in that the accusations may even be correct and on target in the motivating case, but they are frequently aimed at the method, not the way it is used in the case in point. Edwin B. Wilson made a nice comment on this in 1927. He wrote, “It is largely because of lack of knowledge of what statistics is that the person untrained in it trusts himself with a tool quite as dangerous as any he may pick out from the whole armamentarium of scientific methodology.”[3]

The seven pillars I will describe and whose history I will sketch are fine tools that require wise and well-trained hands for effective use. These ideas are not part of Mathematics, nor are they part of Computer Science. They are centrally of Statistics, and I must now confess that while I began by explicitly denying that my goal was to explain what Statistics is, I may by the end of the book have accomplished that goal nonetheless.

I return briefly to one loose end: What exactly does the passage in Proverbs 9:1 mean? It is an odd statement: “Wisdom hath built her house, she hath hewn out her seven pillars.” Why would a house require seven pillars, a seemingly unknown structure in both ancient and modern times? Recent research has shown, I think convincingly, that scholars, including those responsible for the Geneva and King James translations of the Bible, were uninformed on early Sumerian mythology and mistranslated the passage in question in the 1500s. The reference was not to a building structure at all; instead it was to the seven great kingdoms of Mesopotamia before the flood, seven kingdoms in seven cities founded on principles formulated by seven wise men who advised the kings. Wisdom’s house was based upon the principles of these seven sages. A more recent scholar has offered this alternative translation: “Wisdom has built her house, The seven have set its foundations.”[4]

Just so, the seven pillars I offer are the fruit of efforts by many more than seven sages, including some whose names are lost to history, and we will meet a good selection of them in these pages.

CHAPTER 1

· AGGREGATION ·

From Tables and Means to Least Squares

The first pillar, Aggregation, is not only the oldest; it is also the most radical. In the nineteenth century it was referred to as the “combination of observations.” That phrase was meant to convey the idea that there was a gain in information to be had, beyond what the individual values in a data set tell us, by combining them into a statistical summary. In Statistics, a summary can be more than a collection of parts. The sample mean is the example that received the earliest technical focus, but the concept includes other summary presentations, such as weighted means and even the method of least squares, which is at bottom a weighted or adjusted average, adjusting for some of the other characteristics of individual data values.

The taking of a mean of any sort is a rather radical step in an analysis. In doing this, the statistician is discarding information in the data; the individuality of each observation is lost: the order in which the measurements were taken and the differing circumstances in which they were made, including the identity of the observer. In 1874 there was a much-anticipated transit of Venus across the face of the sun, the first since 1769, and many nations sent expeditions to places thought to be favorable for the viewing. Knowing the exact time from the beginning to the end of the transit across the sun could help to accurately determine the dimensions of the solar system. Were numbers reported from different cities really so alike that they could be meaningfully averaged? They were made with different equipment by observers of different skills at the slightly different times the transit occurred at different locations. For that matter, are successive observations of a star position made by a single observer, acutely aware of every tremble and hiccup and distraction, sufficiently alike to be averaged? In ancient and even modern times, too much familiarity with the circumstances of each observation could undermine intentions to combine them. The strong temptation is, and has always been, to select one observation thought to be the best, rather than to corrupt it by averaging with others of suspected lesser value.

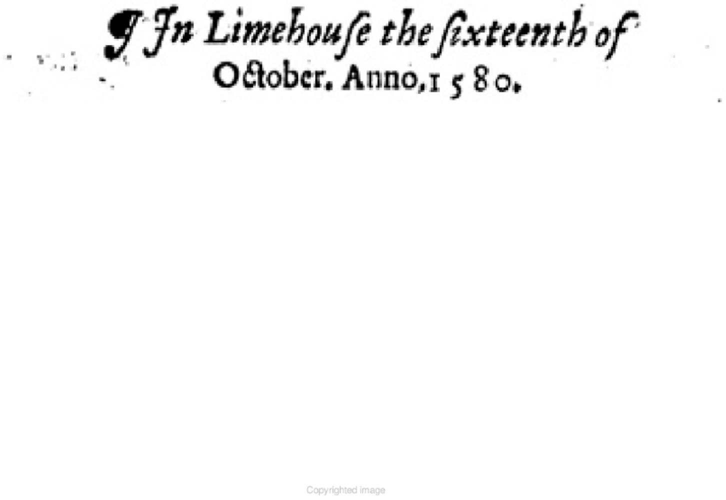
Even after taking means had become commonplace, the thought that discarding information can increase information has not always been an easy sell. When in the 1860s William Stanley Jevons proposed measuring changes in price level by an index number that was essentially an average of the percent changes in different commodities, critics considered it absurd to average data on pig iron and pepper. And once the discourse shifted to individual commodities, those investigators with detailed historical knowledge were tempted to think they could "explain" every movement, every fluctuation, with some story of why that particular event had gone the way it did. Jevons's condemnation of this reasoning in 1869 was forceful: "Were a complete explanation of each fluctuation thus necessary, not only would all inquiry into this subject be hopeless, but the whole of the statistical and social sciences, so far as they depend upon numerical facts, would have to be abandoned."[1] It was not that the stories told about the data were false; it was that they (and the individual peculiarities in the separate observations) had to be pushed into the background. If general tendencies were to be revealed, the observations must be taken as a set; they must be combined.

Jorge Luis Borges understood this. In a fantasy short story published in 1942, "Funes the Memorious," he described a man, Ireneo Funes, who found after an accident that he could remember absolutely everything. He could reconstruct every day in the smallest detail, and he could even later reconstruct the reconstruction, but he was incapable of understanding. Borges wrote, “To think is to forget details, generalize, make abstractions. In the teeming world of Funes there were only details.”[2] Aggregation can yield great gains above the individual components. Funes was big data without Statistics.

When was the arithmetic mean first used to summarize a data set, and when was this practice widely adopted? These are two very different questions. The first may be impossible to answer, for reasons I will discuss later; the answer to the second seems to be sometime in the seventeenth century, but being more precise about the date also seems intrinsically difficult. To better understand the measurement and reporting issues involved, let us look at an interesting example, one that includes what may be the earliest published use of the phrase “arithmetical mean” in this context.

Variations of the Needle

By the year 1500, the magnetic compass or “needle” was firmly established as a basic tool of increasingly adventurous mariners. The needle could give a reading on magnetic north in any place, in any weather. It was already well known a century earlier that magnetic north and true north differed, and by 1500 it was also well known that the difference between true and magnetic north varied from place to place, often by considerable amounts—10° or more to the east or to the west. It was at that time believed this was due to the lack of magnetic attraction by the sea and the consequent bias in the needle toward landmasses and away from seas. The correction needed to find true north from a compass was called the variation of the needle. Some navigational maps of that magnetic north agree at Limehouse, that common value should be (nearly) the midpoint between the two measurements, since the sun travels a symmetrical arc with the maximum at the meridian ("high noon"). On the other hand, if magnetic north is 10° east of true north, then the morning shadow should be 10° farther west and the afternoon shadow likewise. In either case the average of the two should then give the variation of the needle. Borough's table of data for October 16, 1580, is presented in Figure 1.2.



¶ In Limehouſe the ſixteenth of October. Anno. 1580.

1.2 Borough's 1580 data for the variation of the needle at Limehouse, near London.

(Norman 1581)

He had data for nine pairs, taken at elevations from 17° to 25° with the morning variations (given in westward degrees) and afternoon variations (given in eastward degrees, so opposite in sign to the morning, except for the 25° afternoon measure, which was slightly westward). Because of the different signs in the morning and afternoon, the variations in the right-hand column are found as the difference of the variations divided by 2. For the pair taken at sun's elevation 23° we have

$$\begin{aligned}(\mathrm{AM} + \mathrm{PM})/2 &= (34^\circ\ 40' + (-12^\circ\ 0'))/2\\ &= (34^\circ\ 40' - 12^\circ\ 0')/2\\ &= (22^\circ\ 40')/2 = 11^\circ\ 20'.\end{aligned}$$

The nine determinations are in quite good agreement, but they are not identical. How could Borough go about determining a single number to report? In a pre-statistical era, the need to report data was clear, but as there was no agreed upon set of summary methods, there was no need to describe summary methods—indeed, there was no precedent to follow. Borough's answer is simple: referring to the right hand column, he writes, "conferring them all together, I do finde the true variation of the Needle or Cumpas at Lymehouse to be about 11 d. ¼, or 11 d. ⅓, whiche is a poinct of the Cumpas just or a little more." His value of 11 d. 15 m. (11° 15') does not correspond to any modern summary measure—it is smaller than the mean, median, midrange, and mode. It agrees with the value for 22° elevation, and could have been so chosen—but then why also give 11 d. 20 m., the figure for 23° elevation? Or perhaps he rounded to agreement with "one point of the compass," that is, the 11 d. 15 m. distance between each of the 32 points of the compass? Regardless, it is clear Borough did not feel the necessity for a formal compromise. He could take a mean of two values from morning and afternoon at the same

INDEX

Adrain, R., 133
Airy, G. B., 56
al-Biruni, 25, 29
Annals of Eugenics, 107–108, 146–147
Arbuthnot, J., 65–71, 99, 162–163
Aristotle, 150, 152
Average man, 33–35
Avicenna, 150–152

Bartlett, M., 84
Bayes, T., 73–76, 137
Bayesian inference, 73–76, 81–82, 136–139
Bellhouse, D., 51n3
Bellini, L., 116
Berkeley, G., 141–142
Bernard, C., 34–35
Bernoulli, D., 70–71, 80–83
Bernoulli, J., 55
Biometrika, 81, 96, 107
Boethius, 29
Bootstrap, 101–102
Borges, J. L., 15–16, 202
Borough, W., 18–22
Bortkiewicz, L. v., 156–158
Boscovich, R. J., 40–43, 174–176
Bouguer, P., 40
Brahe, Tycho, 22–23, 25
Brahmagupta, 30
Bravais, A., 133
“By the Numbers,” 63–64

Cassini, D., 38
Cauchy distribution, 55
Causal inference, 141–146
Central Limit Theorem, 53–54

Chi-square test, 28, 82–83, 183–186
Clinical trials, 150, 165–166
Composite portraiture, 35–38
Conditional distribution, 132–133, 136–137
Contingency table, 27–28
Correlated observations, 56–57
Cox, D. R., 158, 186–187, 201
Cross-validation, 101
Cushny-Peebles data, 92–95

Dale, A., 137n19
Daniel 1:8–16, 150
Darwin, C., 107–113, 115, 129–131, 173
"D. B.," 25–26
de Finetti, B., 138
Degree of freedom, 183–186
Didion, I., 133–134

Earth, shape of, 38–43, 174–176
Ebbinghaus, H., 160
Edgeworth, F. Y., 59, 97–101, 133, 155–156
Edwards, A. W. F., 83n14
Efron, B., 101–102
Eisenhart, C., 29
Ellipticity, 39, 42–43
Englund, R., 27
Eublides of Miletus, 45
Euclid, 108–111
Euler, L., 167–177

Farr, W., 190–194
Fechner, G., 159
Fienberg, S. E., 165n15
Fisher, R. A., 70, 83, 91, 95–97, 131, 152–155, 161–166, 180, 185–186
Fisher information, 61–62
Fluctuation, 97–100
Friedman, M., 131–132
F-statistic, 101, 163

Galen, 46
Galton, F., 35–37, 87–88, 105, 106, 110–133
"Garden of forking paths," 202
Gauss, C. F., 81–83, 91
Gavarret, J., 77–78
Gellibrand, H., 17, 22
Gelman, A., 202

Gilbert, W., 17
Gosset, W. S., 88–96, 164
Graphical displays, 189
Graunt, J., 108
Guinness Brewery, 88–89, 95
Gunter, E., 22

Harvey, W., 116
Heap, paradox of, 45–47
Herschel, J., 147, 171–173
Herschel, W., 103, 171
Hill, A. B., 143–144, 166
Hume, D., 72–75
Hutton, C., 109
Huygens, C., 66

Ibn Sina, 150–152
Index numbers, 15

Jackknife, 101
Jastrow, J., 160–161
Jeffreys, H., 138
Jenkin, F., 112
Jevons, W. S., 15, 102–106, 152

Kepler, J., 25
Keynes, J. M., 132
Kiaer, A., 164–165
Köbel, J., 31–32
Kruskal, W. H., 165n14

Labby, Z., 181–182, 186
Lagrange, J. L., 80, 135, 176
Lambert, J. H., 80, 82–83
Laplace, P. S., 53–55, 76–77, 81–83, 110, 135, 137–138, 176–177
Lawrence, T. E. (Lawrence of Arabia), 3, 205
Least squares, 43, 82, 132, 141, 176, 180, 189
Legendre, A. M., 43, 82
Lexis, W., 99
Likelihood function, 61, 83–86, 134
Lottery, 166–170

Marshall, A., 202
Master of Kung Fu, 158
Maximum likelihood, 81, 83–85
Mean, 24, 26; variation of, 49–50, 51–52