

Contents

1. THEORIES OF DATA ANALYSIS: FROM MAGICAL THINKING THROUGH CLASSICAL STATISTICS PERSI DIACONIS	1
<p>Exploratory data analysis seeks to reveal structure, or simple descriptions, in data. Any search for structure is a hazardous ritual that can approach magical thinking, unless we allow for the possibility that the revealed structure could have arisen by chance. Even so, exploratory search, when suitably controlled and properly balanced by more classical statistical methods, has been an important contributor to scientific progress. Some emerging theories of data analysis seek to provide a framework for exploratory data analysis as an ingredient of formal scientific inquiry.</p>	
1A. Intuitive Statistics—Some Inferential Problems	4
1B. Multiplicity—A Pervasive Problem	9
1C. Some Remedies	12
1D. Theories for Data Analysis	22
1E. Uses for Mathematics	29
1F. In Defense of Controlled Magical Thinking	31
2. FITTING BY ORGANIZED COMPARISONS: THE SQUARE COMBINING TABLE KATHERINE GODFREY	37

The square combining table offers an alternative resistant fitting strategy for obtaining an additive fit to a two-way data table. Not as tolerant as median polish, it resists very well the impact of a few unusual data values. The basic approach summarizes the differences between each pair of rows and those between each pair of columns,

conveniently using median differences. These medians are then combined, in a way related to usual analyses of paired-comparisons data, to yield row effects and column effects. This approach may be preferable to median polish, particularly when the two-way table contains holes and has at most 20% bad data values.

2A. Combining Comparisons	37
2B. Two-Way Tables	39
2C. Paired Comparisons	47
2D. Analyzing Tables Containing Holes	49
2E. Summary	61
Exercises	62

3. RESISTANT NONADDITIVE FITS FOR TWO-WAY TABLES

JOHN D. EMERSON AND GEORGE Y. WONG

67

Two-way tables of data may exhibit structure that goes beyond that describable in a simple additive fit. One approach to such nonadditive structure (discussed by Mandel and others) adds a general multiplicative term to produce an additive-plus-multiplicative model. A generalization of median polish yields a resistant fit of this model. New resistant strategies supplement existing ones in assessing various fits and their residuals.

3A. The Simple Additive Model and Median Polish	68
3B. One Step Beyond an Additive Fit	71
3C. Assessing and Comparing Fits	79
3D. Multiplicative Fits	83
3E. Techniques for Obtaining Simple Multiplicative Fits	92
3F. Additive-Plus-Multiplicative Fits	100
3G. Some Background for Nonadditive Fits	113
3H. Summary	117
Exercises	119

4. THREE-WAY ANALYSES

NANCY ROMANOWICZ COOK

125

Tables with a measured response and three or more factors extend the ideas of the two-way table. To gain protection against the adverse effects of isolated unusual data values, median polish is extended to higher-way tables. The analyses parallel the more traditional analysis based on means, to which they are compared. Generalizations of the

two-way diagnostic plot aid in detecting systematic patterns of nonadditivity and in learning whether a power transformation would help promote additivity.

4A.	Structure of the Three-Way Table	126
4B.	Decompositions and Models for Three-Way Analysis	128
4C.	Median-Polish Analysis for the Main-Effects-Only Case	130
4D.	Nonadditivity and a Diagnostic Plot in Main-Effects-Only Analysis	145
4E.	Analysis Using Means	158
4F.	Median-Polish Analysis for the Full-Effects Case	164
4G.	Diagnostic Plots for the Full-Effects Case	176
4H.	Fitting the Full-Effects Model by Means	180
4I.	Computation, Other Polishes, and Missing Values	182
4J.	Summary	183
	Exercises	185
5.	IDENTIFYING EXTREME CELLS IN A SIZABLE CONTINGENCY TABLE: PROBABILISTIC AND EXPLORATORY APPROACHES FREDERICK MOSTELLER AND ANITA PARUNAK	189
	<p>An archaeological example illustrates three methods of identifying outliers in the analysis of large contingency tables. The simulation method generates random entries for tables with independent rows and columns and given margins to get the distribution of the largest entries, using a new standardization. The second method uses fixed margins and applies an exploratory approach to locate outliers among the standardized residuals. The third method adjusts the margins to reduce the impact of anomalous cell counts and again applies an exploratory approach to the residuals to locate outliers. The new standardization improves the normal approximation of the far right-hand tail probability for such right-skewed distributions as binomial, Poisson, and hypergeometric.</p>	
5A.	The Hypergeometric Distribution	192
5B.	Assessing Outliers	195
5C.	The Simulation Approach	199
5D.	Applying the Simulation Approach to the Table of Archaeological Data	206
5E.	An Exploratory Approach, Based on Deviations from Independence	212
5F.	A Logarithmic Exploratory Approach	214

5G.	Illustrations of the New Standardization	217
5H.	Summary	221
5I.	Conclusion	223
6.	FITTING STRAIGHT LINES BY EYE	
	FREDERICK MOSTELLER, ANDREW F. SIEGEL, EDWARD TRAPIDO, AND CLEO YOUTZ	225
	<p>Because investigators frequently fit lines by eye, it is useful to know something about the properties of such a procedure. Students fitted lines by eye to each of four sets of points presented in an experimental design. Their pooled slope was closer to the slope of the major axis than to the slope of the least-squares regression line, and the median efficiency of fitting was about 63 percent.</p>	
6A.	Method	226
6B.	Results	229
6C.	Summary	238
7.	RESISTANT MULTIPLE REGRESSION, ONE VARIABLE AT A TIME	
	JOHN D. EMERSON AND DAVID C. HOAGLIN	241
	<p>When regression data come with more than one carrier (or predictor), repeated application of the exploratory resistant line guards against the effects of isolated wild data values. Taking the carriers in a specified order, the analysis sweeps the effect of each carrier out of the response and out of all later carriers. A related approach provides a resistant analog of two-way analysis of covariance.</p>	
7A.	Resistant Lines	242
7B.	Sweeping Out	246
7C.	Example	250
7D.	When Carriers Come in Blocks	263
7E.	Summary	273
	Exercises	275
8.	ROBUST REGRESSION	
	GUOYING LI	281
	<p>Robust regression methods, besides providing estimates that are often more useful, can call attention to unusual data in a regression data set.</p>	

These methods, including those based on M -estimators and W -estimators as well as bounded-influence regression, protect against distortion by anomalous data and have good efficiency over a wide range of possibilities for the error structure.

8A.	Why Robust Regression?	282
8B.	M -Estimators and W -Estimators for Regression	291
8C.	Computation	304
8D.	Example: The Stack Loss Data	310
8E.	Bounded-Influence Regression	322
8F.	Some Alternative Methods	328
8G.	Summary	335
	Exercises	341
9.	CHECKING THE SHAPE OF DISCRETE DISTRIBUTIONS	
	DAVID C. HOAGLIN AND JOHN W. TUKEY	345
	<p>New graphical techniques help to assess how closely an observed discrete frequency distribution follows a member of one of the common families of discrete distributions, such as the binomial or the Poisson. Used carefully, these techniques prevent unusual counts from distorting the overall assessment. The slope of a straight-line pattern in the plot identifies the main parameter of the family of distributions. Other plots help to show whether a cell is discrepant, as well as to assess whether different parts of the data indicate different values of the main parameter.</p>	
9A.	A Poissonness Plot	348
9B.	Confidence Intervals for the Count Metameter	358
9C.	When Is a Point Discrepant?	370
9D.	Overall Plots for Other Families of Distributions	376
9E.	Frequency-Ratio Alternatives	389
9F.	Cooperative Diversity	396
9G.	Double-Root Residuals	406
9H.	Summary	409
	Exercises	412
10.	USING QUANTILES TO STUDY SHAPE	
	DAVID C. HOAGLIN	417

Instead of the classical measures based on the third and fourth moments, numerical and graphical techniques based on quantiles

serve to describe skewness and elongation. Variants of quantile-quantile plots aid in comparing distributions.

10A.	Diagnosing Skewness	419
10B.	Diagnosing Elongation	425
10C.	Quantile-Quantile Plots	432
10D.	Plots for Skewness and Elongation	442
10E.	Pushback Analysis	450
10F.	Summary	454
10G.	Appendix	456
	Exercises	459

11. SUMMARIZING SHAPE NUMERICALLY: THE g -AND- h DISTRIBUTIONS

DAVID C. HOAGLIN

461

This chapter approaches the study of distribution shape more quantitatively. Selected monotonic functions of a standard Gaussian random variable can be described by constants g and h that indicate skewness and elongation, respectively. The resulting family of g -and- h distributions offers resistance and flexibility in summarizing distribution shapes quantitatively.

11A.	Skewness	462
11B.	Elongation	479
11C.	Combining Skewness and Elongation	485
11D.	More General Patterns of Skewness and Elongation	490
11E.	Working from Frequency Distributions	496
11F.	Moments	501
11G.	Other Approaches to Shape	504
11H.	Summary	508
	Exercises	511

INDEX

515