

Introduction to Statistical Methods in Pathology

Amir Momeni
Matthew Pincus
Jenny Libien

 Springer

Introduction to Statistical Methods in Pathology

Amir Momeni • Matthew Pincus
Jenny Libien

Introduction to Statistical Methods in Pathology

 Springer

Amir Momeni
Department of Pathology
State University of New York
Downstate Medical Center
Brooklyn, New York, USA

Matthew Pincus
Department of Pathology
State University of New York
Downstate Medical Center
Brooklyn, New York, USA

Jenny Libien
Department of Pathology
State University of New York
Downstate Medical Center
Brooklyn, New York, USA

ISBN 978-3-319-60542-5 ISBN 978-3-319-60543-2 (eBook)
DOI 10.1007/978-3-319-60543-2

Library of Congress Control Number: 2017944203

© Springer International Publishing AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

To an ever-increasing extent, pathologists are being required to use statistics in their practice. In clinical pathology or laboratory medicine, statistics are a fundamental requirement for the evaluation of the reliability of quantitative results for values of serum and, in general, body fluid analytes such as electrolytes, glucose, blood urea nitrogen (BUN), creatinine, critical enzymes, etc. and for the analysis of the correlation between the results generated on different analyzers, all of which are used for quantitative determination of the same analytes. Correlations between measurements of parameters that allow categorization of tumors such as correlation of nuclear grade with pathological stage require knowledge of statistical methods in anatomic pathology. Correlation of the staging of different cancers with survival involves another major use of statistics in both anatomic and clinical pathology.

Often, pathologists utilize statistical methods without knowledge of the physical and mathematical basis that underlies the particular statistics that they are using. This can give rise to erroneous conclusions. For example, many, but certainly not all, quantitative analyses for analytes follow so-called Gaussian statistics, an example of parametric statistics, with a known mathematical form for the distribution of values that gives rise to the “bell-shaped curve,” called the normal distribution. This involves computation of means, standard deviations, confidence intervals for means, and a number of other parameters.

However, these methods cannot be used for analyte values that do not follow Gaussian statistics which requires that the distribution of values for a given analyte be distributed in what is termed a normal distribution as represented by the so-called bell-shaped curve. This can affect determinations such as the reference ranges for analytes based on values determined from presumed normal or well individuals. If the distribution of values is assumed to be Gaussian, the range would be computed as the mean of the values plus or minus two standard deviations from the mean. However, if the values actually do not follow a Gaussian distribution, serious errors can be made in establishing the reference range which may be too narrow or too wide. Not infrequently, the use of non-parametric statistics rather must be used in establishing reference ranges.

We are currently living in what has been termed “the age of metrology.” This means that, to an increasing degree, statistics govern most aspects of laboratory medicine including whether or not values can be accepted as being “true” or

“reliable,” and the criteria for acceptability are being made more stringent. This raises the question as to why statistics are considered essential in evaluation of clinical laboratory results.

Statistics provide a means for at least partially removing arbitrariness for making such critical decisions as whether results are acceptable or whether two sets of data are actually the same or are different. However, there are limitations to statistical analysis.

In all statistical analysis, there is some arbitrariness. For example, analysis of the concentration of an analyte in a control sample is said to be “acceptable” if the value lies between plus or minus two standard deviations of the mean determined for concentration of this analyte in the control sample on a clinical chemistry analyzer. The reason for this two standard deviation rule is that, for a Gaussian or normal distribution, the mean plus or minus two standard deviations encompass about 95% of the possible values. All other values are considered to be “outliers.” This is an arbitrary number. One can inquire why some other number might be used such as 97% (allowed by using approximately three standard deviations from the mean) or some other number. Here, there is no definitive answer.

Given the current metrology requirements and their acceptance by federal and state regulatory agencies and by most laboratorians and given the necessity for use of statistics in analyzing specific clinical data, it is desirable to introduce pathologists to the statistical methods available to them so that they understand what methods to use in analyzing clinical data and how to use them. It is the purpose of this textbook to achieve this goal.

Our aim, therefore, is to impart to the reader how to evaluate different types of data using the appropriate statistical methods and why these methods are used, rather than to refer the reader to specific programs that analyze the data without explanation of the basis of the methods used. In this textbook, we present the most commonly used statistical methods in the field of pathology. Our presentation is based on three simple steps:

1. *Definition of the statistical problem.* For example, when a control is assayed, the statistical problem is to determine whether the result is acceptable or not acceptable.
2. *The mathematical form of the statistical distribution that solves the statistical problem.* Using the same example given above, since the assay is performed on the same control repeatedly, any deviation of the values from one another should be random, i.e., there is random error. Random error is described by the Gaussian distribution, i.e., when the probability of getting a particular value is plotted against the values themselves, a bell-shaped curve is obtained. The mathematical form for this bell-shaped probability distribution is the exponential form ae^{-bx^2} , where x is any value determined experimentally and a and b are constants related to the standard deviation.
3. *How to compute the significance of results obtained from data obtained in the medical laboratory using the appropriate distribution.* The mean for the Gaussian distribution can be shown to be the most probable value on the bell-shaped

curve and equals the median value. From the Gaussian distribution, one standard deviation from the mean can be computed. It can further be shown that approximately 95% of all values lie within the width of the bell-shaped curve at two standard deviations. It happens that one standard deviation can also be computed as the square root of the sum of the squares of the differences between each value determined experimentally and the mean value divided by the number of values.

Thus, as we discussed above, if we wish to define acceptability of a value as any value that lies within two standard deviations of the mean value, then if a result is within this cutoff, i.e., plus or minus two standard deviations from the mean value, it is acceptable.

The textbook is arranged so that the most commonly used statistics in pathology are discussed first in Chap. 2 in which normal or Gaussian distributions are described; the concepts of accuracy and precision are discussed; the evaluation of test efficacy, i.e., sensitivity, specificity, and positive and negative predictive value, is presented; and the evaluation of so-called receiver operator curves is performed in deciding which of two or more tests has better or best diagnostic accuracy.

Chapter 3 then presents general probability analyses and discusses probability distributions that are not used as frequently in pathology but may be useful, especially the ones involving conditional probabilities.

Chapter 4 presents the underlying theory for analyzing correlations, e.g., when samples are analyzed on two or more analyzers, what criteria are to decide whether the values obtained on assaying samples can be considered to be the same or different. This chapter discusses how to fit straight lines to experimentally determined points, a process termed linear regression analysis, and how to decide how well the “best fit” line fits the points.

Chapter 5 provides the statistical basis for the all-important question as to whether a new test for diagnosis of a particular disease is valid. Generally, these tests provide a yes or no answer, i.e., the results are discrete and not continuous. It happens that the distribution that is most appropriate for answering this question of reliability of the test is given by the chi-squared distribution. A major point of this chapter is to illustrate that, although the results of this type of testing are discrete, it is possible to represent the probability distribution for right or wrong results as a continuous function so that cutoffs such as those used for the Gaussian distribution can be used and quantitative decisions can therefore be made.

Chapter 6 addresses the statistical basis for the comparison of two or more sets of data to determine whether they are the same or different. This involves specific tests on the mean values for the sets of data.

Chapter 7 discusses multivariate analysis, i.e., extension of the linear regression analysis discussed in Chap. 4 to linear regression with more than two variables.

Chapter 8 presents methods for inferring values omitted from datasets that are necessary for statistical analysis.

Chapter 9 presents the statistical solution to a problem that is common to all medical practice: survival analysis. Many readers may be familiar with Kaplan-Meier curves for survival of patients who carry specific diagnoses or who are being

treated for specific diseases. This chapter explains the statistical basis for this type of analysis and other approaches that achieve the same goal.

Chapters 10 and 11 deal with quality assurance. Chap. 10 addresses how methods are quantitatively validated and Chap. 11 discusses the rules for evaluating quality control.

Chapters 12 and 13 deal with the problems of how to evaluate quantitatively and how to design diagnostic studies.

Chapter 14 is an introduction to statistical analysis of large datasets. This type of analysis is now becoming of paramount importance as the amount of genetic information on patients has been increasing exponentially. In this chapter, the technique of clustering, which allows for data simplification, is discussed.

We hope that the readers of this textbook will find it helpful to them in evaluating data in clinical practice and/or in research.

Brooklyn, NY, USA

Matthew R. Pincus
Amir Momeni
Jenny Libien

Contents

1	Why Every Pathologist Needs to Know Statistics	1
	Introduction	1
	Outcomes and Variables in Pathology and Laboratory Medicine	1
	Components of a Useful Diagnostic Test	3
	Examples	3
	Defining Test Objectives	4
	Summary	5
	References	5
2	Assessing Diagnostic Tests	7
	Technical Accuracy and Precision	8
	Error	8
	Standard Deviation	11
	Confidence Interval	12
	Calculating Reference Intervals	14
	Calculating Sample Size for Reference Interval Estimation	17
	Diagnostic Accuracy and Testing for Accuracy	18
	Sensitivity and Specificity	19
	Predictive Values	20
	Receiver Operating Characteristic Curve	23
	Calculating AUC	26
	Clinical Applicability	28
	Transferability	30
	Feasibility	31
	Cost-Effectiveness Analysis	31
	Summary	33
	References	36
3	Probability and Probability Distribution	39
	Introduction	39
	Probability Measure and Axioms of Probability	42
	Conditional Probability	44
	Multiplication Rule	46

Bayesian Probability	47
Pretest and Posttest Probability	49
Probability Distribution	52
Discrete Distribution	53
Mean and Variance	60
Continuous Distributions	63
Introduction to Distribution Plots	68
Summary	71
References	72
4 Linear Correlations	75
Linear Correlations in the Medical Laboratory	75
Two-Tailed T-Test	75
Correlation Plots	76
Determination of the “Best” Straight Line Through Experimentally Determined Points	78
Derivation of the Least Square Best Fit Line Through the Experimentally Determined Points	79
Correlation Coefficient	81
Problems with This Approach	82
Slopes and Intercepts	83
Errors in the Slopes and Intercepts	83
Error in the Slope	84
Error in the Intercept (S_{int})	86
Bias	87
Linearity and Calibration	87
Practical Considerations	89
Calibration	89
Summary	90
References	91
5 Cross Tabulation and Categorical Data Analysis	93
Introduction	93
Categorical Variables	93
Contingency Table	94
Hypothesis Testing	97
Analysis of Risk Ratios	102
Chi-Squared Tests	103
Degrees of Freedom	105
Chi-Squared Distribution	105
Pearson Chi-Squared Test	107
McNemar’s Test	111
Cochran–Mantel–Haenszel Test	112
Fisher’s Exact Test	114
Measures of Agreement	116

Cohen’s Kappa	117
Fleiss’s Kappa	118
Summary	119
References	120
6 Comparing Sample Means	121
Introduction	121
Continuous Data	121
Mean and Median	122
Variance, Skewness, and Kurtosis	123
Parametric Versus Non-parametric Tests	124
Outliers	125
One-Tailed Versus Two-Tailed Testing	126
Testing for Normality	128
Parametric Tests	133
Student’s t-Test	134
One-Way ANOVA	142
Non-parametric Tests	146
Mann-Whitney U Test	147
Kruskal-Wallis Test	150
Effect Size	151
Cohen’s <i>d</i>	151
Cohen’s <i>f</i>	152
Ordinal Variables	152
Kendall’s Tau Test	153
Spearman’s Rho Test	154
Summary	156
References	157
7 Multivariate Analysis	159
Introduction	159
Generalized Linear Model	160
Multiple Regression Analysis	161
Assessing Utility of the Fitted Model	163
Interaction and Collinearity	167
Logistic Regression	168
Binary Logistic Regression	168
Multinomial Logistic Regression	174
Ordinal Logistic Regression	177
Internal and External Validity	180
Summary	184
References	184
8 Imputation and Missing Data	185
Introduction	185
Missing Data	185

Types of Missing Data	186
Graphical Visualization of Missing Data	190
Dealing with Missing Data	191
Robust Statistics	192
Data Discarding Solutions	192
Complete-Case Analysis	193
Available-Case Analysis	193
Imputation	194
Single Imputation	194
Multiple Imputation	196
Summary	200
References	200
9 Survival Analysis	201
Introduction	201
Incidence	201
Survival Analysis	203
Censoring	204
Survival Data	204
Survival Function	205
Hazard Function	205
Kaplan-Meier Estimator	208
Log-Rank Test	211
Cox-Proportional Hazards Regression	214
Summary	217
References	217
10 Validation of New Tests	219
Introduction	219
Test Validation	220
Defining Analytical Goals	221
Validation Experiments	223
Sample Size Calculations	224
Accuracy Experiment for Qualitative Tests	226
Precision Experiment for Qualitative Tests	226
Method Comparison Experiments for Quantitative Tests	227
F-Test for Precision	232
Linearity Experiments for Reportable Range	233
Allowable Total Error	236
Detection Limit Experiments	238
Notes on Validation of Immunohistochemical Tests	239
Summary	241
References	241

11	Statistical Concepts in Laboratory Quality Control	243
	Introduction	243
	Control Limits	244
	Levey-Jennings Charts	245
	Westgard Rules	246
	Average of Normals	250
	Delta Check	251
	Moving Patient Averages	253
	Statistical Concepts for External Quality Control	255
	Summary	256
	References	256
12	Critical Appraisal of Diagnostic Studies	259
	Introduction	259
	Levels of Evidence	260
	Evidence-Based Recommendations	262
	Critical Appraisal of Diagnostic Studies	264
	Systematic Reviews	266
	Meta-analysis	269
	Publication Bias	275
	Summary	277
	References	277
13	Designing Diagnostic Studies	279
	Introduction	279
	Diagnostic Research Design	279
	Phases in Clinical Diagnostic Studies	282
	Diagnostic Accuracy Studies	284
	Index Test	285
	Reference Standards	285
	Examples of Diagnostic Accuracy Study Designs	287
	Observational Studies	288
	Paired Comparative Accuracy Studies	289
	Randomized Comparative Accuracy Studies	290
	Sample Size Calculations	290
	Reporting of Diagnostic Accuracy Studies	291
	Summary	292
	References	292
14	Statistical Concepts in Modern Pathology Practice	293
	Introduction	293
	Clustering Algorithms	294
	<i>K</i> -Means Clustering	294
	Hierarchical Clustering	296
	Summary	299
	References	299

Appendix A	301
Appendix B	303
Appendix C	305
Appendix D	307
Index	311

Introduction

Statistics permeates our lives as pathologists. We use statistics in the interpretation of laboratory tests, in deciding whether to use a new immunohistochemical stain or diagnostic method, for our research projects, in our critical reading of scientific literature, and in our quality improvement and laboratory management activities [1]. Although we regularly use statistics as pathologists, we do not understand statistics as well as we would like. In a survey of pathologists to assess statistical literacy (Schmidt et al., Arch Pathol Lab Med, 2016) [2], the majority of pathologists surveyed expressed the desire to have a better understanding of statistics. This book aims to help pathologists achieve a higher level of statistical literacy and gain greater comfort in using statistical methods.

We start now with the basics – the definitions of keywords. The Merriam Webster definition of statistics is “a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data” or “a collection of quantitative data.” For pathologists, statistics can be thought of as a way to make sense of our observations and measurements.

Outcomes and Variables in Pathology and Laboratory Medicine

The test result can have different formats. We can consider test results as variables. ‘Variables’ are data items which can be counted or measured. Variables may be categorical, which is also known as nominal, and have a limited number of data items such as positive staining vs. negative staining for a surgical pathology project examining the use of a new immunohistochemical stain. Alternatively, variables may be ordinal in which the variables are ordered, but there is still a limited number of data items (but often more than in categorical). An example of an ordinal scale could also use that same surgical pathology project examining the use of a new immunohistochemical stain; however, a semiquantitative assessment is performed,