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

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^{-bx2}, 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