

Biostatistics 101: Understanding Data

David Etzioni, MD, MSHS
Maher A Abbas, MD, FACS, FASCRS

Introduction

Kaiser Permanente (KP) is a leader in health care delivery and provides care for millions of Americans in several regions and states, including Northern and Southern California, Colorado, Georgia, Hawaii, Ohio, the Northwest, and the Mid-Atlantic States. The volume of clinical care rendered every year throughout the organization presents a great opportunity for research and innovations. In recognition of the importance of research, the Kaiser Foundation Research Institute was created in 1958 to administer and support research within KP at a national and regional level. High-quality innovative translational research is performed every year, in the form of randomized clinical studies, epidemiologic research, retrospective databases review, and health care policy research. Supported by an electronic medical record and computerized databases, KP is well-suited to provide the scientific community with a wealth of data on outcome of interventions. In 2007, there were approximately 2800 active studies, all approved by institutional review boards, being conducted nationally. That same year, reports on 571 studies were published by KP physicians and scientists in prestigious medical, surgical, and scientific journals, including *The Permanente Journal*.

Because of the size of the KP population, KP research studies often contain a large number of study subjects. Such projects and endeavors can generate complex data and results that require analysis to demonstrate the effect of therapies and interventions, to establish the efficacy or limitation of treatments, and to prove or to refute a scientific hypothesis. An understanding of biostatistics is critical to the researcher investigating clinical questions. Equally important is an appreciation of statistics by the reader and interpreter of published studies. As with all fields of scientific endeavor, statistics encompasses a rich jargon that is necessary to abbreviate and refer to underlying concepts. In this article, the first of a three-part series on statistics for clinicians, we begin with an overview of how statistics can and should be used to describe data.

Types of Data

Data are available in a wide range of types, and understanding the type of data at hand is a crucial first step in any statistical analysis. Broadly speaking, data can be quantitative or qualitative. Quantitative data are reported in units of measurement. There are two main categories of quantitative data:

- *Continuous data* represent measurements on a spectrum, where a data element can have any one of an infinite number of intermediate values. Age is an example of this. Whereas age is generally reported in years—for example, 45 years of age—there is no reason why it could not be reported in tenths or hundredths of years, such as 45.3 years or 45.36 years.
- *Discrete data* represent whole-number measurements that cannot be split—for example, number of children or number of previous hospitalizations. Clearly the number might vary widely from person to person, but there is no possibility for a value that is not a whole number.

Qualitative data are reported in categories. As with discrete data, there is no opportunity for an intervening value. The main difference between qualitative data and continuous data is that every data element has a value from a preconceived list of possibilities. There are two main categories of qualitative data:

- *Ordered data* represent measurements along a spectrum—for example, a visual pain scale with options of “no pain,” “mild pain,” “moderate pain,” and “severe pain.”
- *Discrete data* represent mutually exclusive options that do not occur along a spectrum. Race/ethnicity is a common example of discrete data.

Describing Data

Different types of data are described differently. Qualitative data can be reported adequately with a report of frequency, in a simple table. Reporting quantitative data is somewhat more complicated. In describing

David Etzioni, MD, MSHS, is an Assistant Professor of Colorectal Surgery and Preventive Medicine at the Keck School of Medicine, University of Southern California. E-mail: etzioni@usc.edu.

Maher A Abbas, MD, FACS, FASCRS, is an Assistant Clinical Professor of Surgery at the University of California, Los Angeles; the Chief, Colon and Rectal Surgery and Chair, Center for Minimally Invasive Surgery at the Los Angeles Medical Center in Los Angeles, CA. E-mail: maher.a.abbas@kp.org.

a set of quantitative observations, we generally rely on measurements of the *center* and the *distribution* of the data.

The *center* of a set of quantitative data is usually reported in one of two ways—the mean (average) value or the median. The median value of a data set is simply the middle value of the list of measurements when it is ordered from least to greatest.

The distribution of a data set is described according to the spread of its values, and several terms are used toward this end. Variance and standard deviation are related terms that measure the amount by which observations in a data set differ from the central (mean) value of the data set. Standard deviation is simply the square root of variance.

In addition to quantifying the distribution of values in a data set of quantitative, continuous measurement, it is also important to know the *character* of the distributions. Initially, this is best done using a histogram. In an example of 1000 measurements of height taken from a fictional educational institution—Biostats Junior High—the mean (average) height is 62.1 inches and the standard deviation is 3.0 inches (Figure 1).

The shape of this curve is common among data sets of continuous measurements and is often referred to as a normal or “Gaussian” distribution.

Knowing that this data set of 1000 observations has this classic bell curve empowers researchers to use a broad range of statistical techniques that assume that this type of distribution is present. One example of such a technique is a convenient rule of thumb regarding standard deviation (SD). In general, for normally distributed data sets, two-thirds of observations will occur within a range that is encompassed by the mean ± 1 SD and 95% of observations will occur within the mean ± 2 SD.

Now, what if the underlying distribution of the data is not normal? In another example—Statistics Summer Camp, which has 700 junior high school students and also a group of 300 younger students—the mean height is 55.8 inches, the SD is 9.9 inches, and the histogram of height looks like Figure 2.

Will our rule of thumb still hold? Within the range encompassed by the mean ± 1 SD, 634 students (63.4%) are fairly close to our two-thirds rule. A total of 978 (97.8%) observations are within the range of the mean ± 2 SD. Despite the fact that our data set is markedly non-normal, our rule of thumb still holds up fairly well. In these situations, we say that our rule is (moderately) *robust* regarding the assumption of normality.

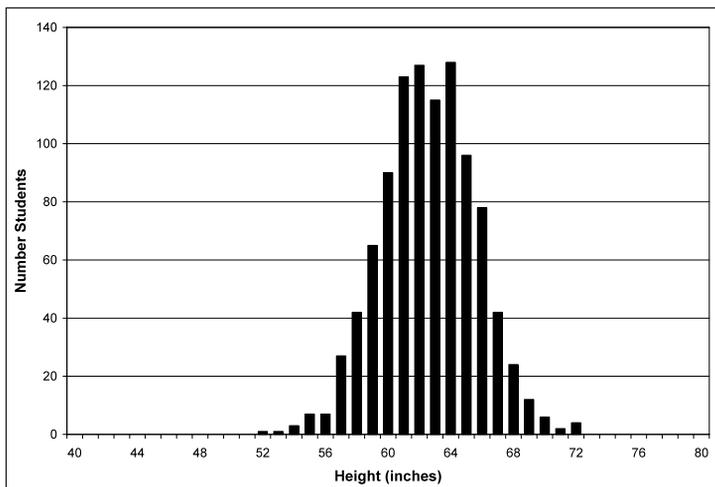


Figure 1. Normal (“Gaussian”) distribution on a histogram.

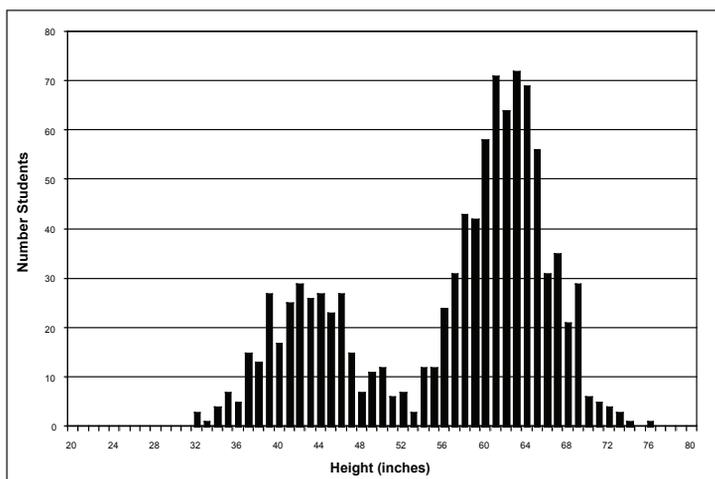


Figure 2. Non-normal distribution on a histogram.

The concept of robustness is important when choosing statistical tests. Different testing techniques rely to differing extents on specific assumptions. In general, statistical tests all have a tradeoff between power (the ability to detect a difference when one is present) and reliance on assumptions.

In the second article in this series, we will explore issues related to the concept of a sample, plus significance testing between two groups. ❖

Disclosure Statement

The author(s) have no conflicts of interest to disclose.

Acknowledgment

Katharine O’Moore-Klopf, ELS, of KOK Edit provided editorial assistance.