A Statistics Primer

Article *in* The American Journal of Sports Medicine · January 1998

DOI: 10.1177/036354659602400324 · Source: PubMed

3 authors, including:

Mary Lou Greenfield
University of Michigan
85 PUBLICATIONS **2,550** CITATIONS

0363-5465/97/2525-0022$02.00/0
THE AMERICAN JOURNAL OF SPORTS MEDICINE, Vol. 25, No. 5

Current Concepts

A Statistics Primer

Descriptive Measures for Continuous Data

Mary Lou V. H. Greenfield,* MPH, MS, John E. Kuhn, MS, MD, and Edward M. Wojtys, MD

From the University of Michigan, Department of Orthopaedic Surgery, Ann Arbor, Michigan

Whether data are collected as part of a research study or to help the physician characterize a group of patients, the resulting data fall into either one of two categories: discrete or continuous. If the outcome of interest includes values that fall into discrete categories (e.g., gender, race, or presence or absence of disease), statistical tests for discrete data are used. (Descriptive measures and a review of statistical tests for these measures were discussed in a previous "Current Concepts" article.[5]) Discrete data are measured in frequencies of occurrence. For example, in a study of injuries among high school athletes, 76 women experienced an injury compared with 24 men. Gender is an example of a discrete variable. In contrast, a continuous variable results in data that can be "characterized by having an infinite number of evenly spaced potential values between any two values."[4] For example, weight is an example of a continuous variable. That is, between 180 and 181 pounds there are an unlimited number of possible units of weight that might be measured in infinitely small increments. Other examples of continuous variables include blood pressure, temperature, and height.

We use descriptive measures to summarize continuous data that portray aspects of our everyday lives. We may discuss that the average rainfall in our state during the months of June, July, and August is 5 inches per month plus or minus 1.2 inches; we may even boast that this rainfall of 5 inches per month is in contrast to the drought in 1978 where the all-time low rainfall was 5 inches for the entire summer, or the flood of 1961 when 11.5 inches fell in one 24-hour period. These precipitation facts are quantitative stories that help us to describe how wet (or dry) our immediate world is. Rainfall data accrued throughout the years provide summary measures with which to compare rainfall in one particular year. Specifically, these descriptive measures relate to the central tendency, the variation, and the distribution of our data. From this weather description we learn about the mean rainfall for this year and other years, and the variation over the years; using these two measures we can create a picture of both the distribution of the rainfall over past years and expected precipitation in future years.

The building blocks of statistical tests for continuous data are precisely these measures: central tendency, variation, and distribution of the data. They are essential to understanding statistical tests such as the Student's *t*-test and analysis of variance (ANOVA) to be discussed in a future "Current Concepts" article.

DISTRIBUTION OF THE DATA

In many studies, investigators are interested in whether a treatment or intervention in a sample being studied can be used to infer information about a population from which the study group was drawn. For example, suppose an investigator wants to know the number of injuries experienced by all United States high school football players during 2 years of varsity football training and games. Realistically, he does not observe the entire distribution of injury rates of all high school football players in the country (the population). Instead, the investigator's best guess of the mean injury rate of this population is generally based on a single sample of players. In this example, the investigator may select a sample of 25 players from a nearby high school and record the number of injuries each player experienced during 2 years of play. Again, summarizing the data from these 25 players will allow the investigator to record the number of injuries in this particular group to draw some inference about the population of all United States high school football players. (Clearly, it is essential that the football players are chosen in such a way that they represent the population of all high school football players; this is because the investigator wants the

* Address correspondence and reprint requests to Mary Lou V. H. Greenfield, MPH, MS, University of Michigan, Orthopaedic Surgery, TC2914G-0328, 1500 East Medical Center Drive, Ann Arbor, MI 48109.
No author or related institution has received any financial benefit from research in this study.

findings from this sample to be a reasonable estimate of the rest of the high school football player population.[3]) "It is important to note that populations are unique, but that samples are not."[1]

Consequently, for all high school football players with 2 years of varsity experience there is only one range of injuries experienced. However, if another investigator drew a sample of football players and recorded the range of injuries for her sample, it would be a different sample than that of the first investigator. Statistical theory has measurement tools for this sample-to-sample variation and uses this information in statistical hypothesis testing. Note the number of injuries in the sample in Table 1. Figure 1 shows the distribution of injuries in the sample.

This reasonably symmetric, bell-shaped distribution of injuries demonstrates the plot or curve of injuries in the sample. This distribution of injuries in the sample approximates the *Normal* or *Gaussian distribution.* It is important to note that the word *Normal* used to describe a bell-shaped distribution does not mean *usual, typical, physiologic,* or *most common;* likewise, a distribution that is non-Normal, is *not abnormal.*[6] The Normal distribution is an abstract pattern approximated by many variables in which the data are bell-shaped, with a single peak at the mean (Fig. 2).

The Normal distribution is fundamental to performing statistical analyses. It is also important because the distribution of many medical measurements in populations are approximately represented by a curve that is Normal in shape, e.g., blood pressure, height, and weight.[2] We assume that the underlying population of means from which the mean for our sample has been drawn has the Normal distribution. The mean and the standard deviation from our sample are used to estimate the population mean and the standard error. The mean is an example of an average and is a measure of central tendency, while the standard deviation is a measure of the spread of the data. (The mean and the standard deviation will be discussed in greater detail later in this article.)

As can be seen in Figure 3, as the mean of a variable being measured changes, the distribution moves along the axis. Figure 4 demonstrates that if the standard deviation is small, most of the data will be centered closely to the mean; if the standard deviation is large, the data will be distributed widely along the axis. It is also important to note that a property of the Normal distribution is that exactly 95% of the values of the distribution lie within 1.96 standard deviations of the population mean.

Occasionally data in a sample will be skewed to one side of the mean with a longer tail on one end (Fig. 5); in such circumstances the data may be transformed using the logarithm or square root of the values, to approximate the Normal distribution. This transformation is beyond the scope of this article.

TABLE 1
Numbers of Injuries Among Players

Injuries	Players
0	0
1	1
2	2
3	3
4	5
5	6
6	3
7	2
8	2
9	1

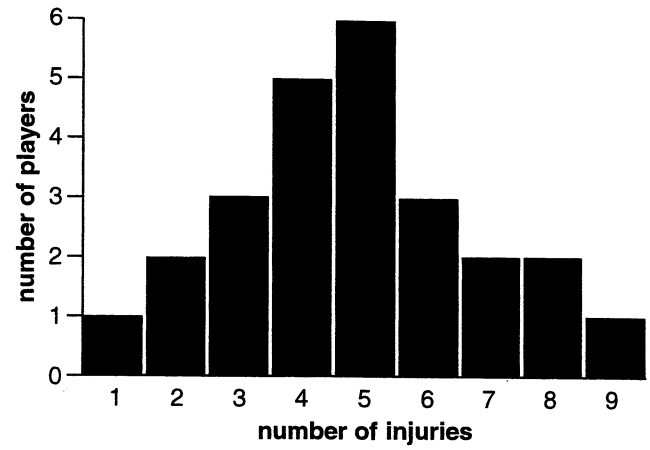

Figure 1. Distribution of injuries among players.

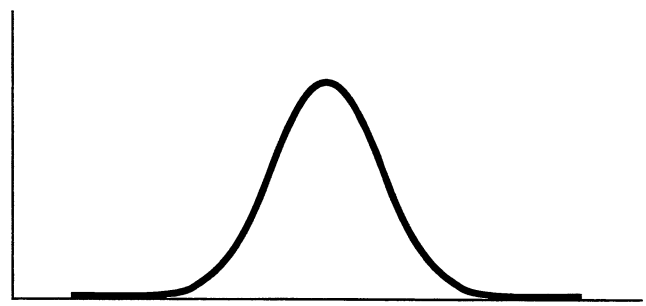

Figure 2. The normal distribution: bell-shaped curve, single peak at the mean, symmetric tails on either side of the mean, 95% of values lie within 1.96 standard deviations from the mean.

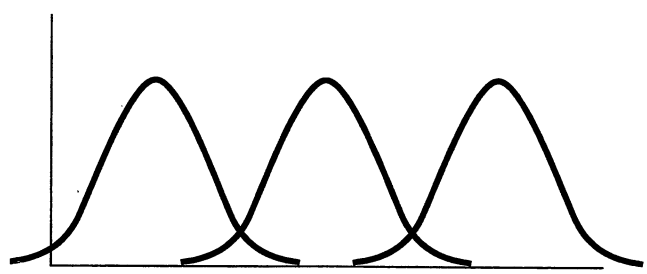

Figure 3. Normal distribution with changing means, symmetric tails.

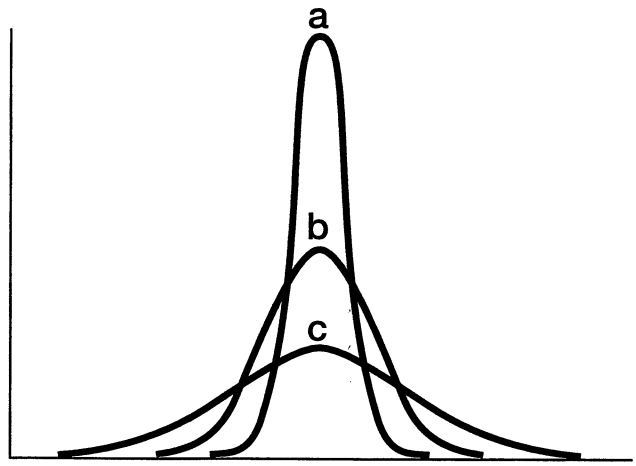

Figure 4. Normal distribution with the same mean but different standard deviations. a, small standard deviation; b, moderate standard deviation; c, large standard deviation.

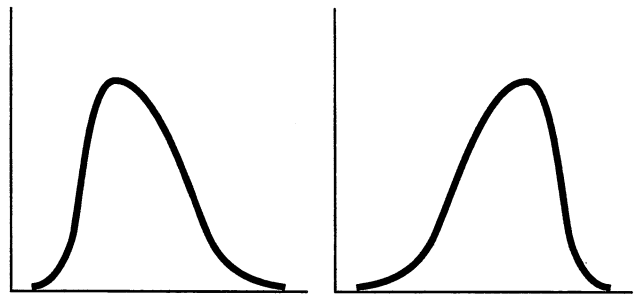

Figure 5. An example of skewed distribution.

TABLE 2
Summary Statistics for a Sample of Weights of 25 Ballplayers[a]

Weight	Mean weight 185.64 lb	Deviation from the average weight (the individual player's weight minus the mean)	Squared deviations
180		−5.64	31.81
230		44.36	1967.81
140		−45.64	2083.01
150		−35.64	1270.21
160		−25.64	657.41
165		−20.64	426.01
170		−15.64	244.61
166		−19.64	385.73
200		14.36	206.21
190		4.36	19.01
195		9.36	87.61
185		−0.64	0.41
187		1.36	1.85
200		14.36	206.21
210		24.36	593.41
177		−8.64	74.65
189		3.36	11.29
195		9.36	87.61
199		13.36	178.49
201		15.36	235.93
178		−7.64	58.37
195		9.36	87.61
190		4.36	19.01
200		14.36	206.21
189		3.36	11.29
			Σ 9151.76

[a] Median = 189; mode = 200; range = 90.
Sample variance (sum of squared deviations/N − 1) $= \frac{9151.76}{24} = 381.323$.
Standard deviation (square root of the sample variance) $= \sqrt{381.32} = 19.52$.

MEASURES OF CENTRAL TENDENCY

Consider the following example. Suppose a coach wanted to describe the weights of a group of 25 ballplayers, as depicted in Table 2. The first thing the coach might look at is the *mean* weight of the ballplayers. The mean is the result of adding the weights of all the players and then dividing by that total by the number of ballplayers in the sample or n. In this example the mean weight of the ballplayers is 185.64 pounds. The mean is one example of a measure of central tendency. Measures of central tendency demonstrate where data in the sample are located or are "centered" along a continuum of possible values. The mean is the most commonly used measure of central tendency. Unfortunately, the mean will be affected by extreme values if the distribution is skewed; for example, what if one of the players weighed 324 pounds? His weight would increase the mean of the sample.

An alternative central tendency measure that is less sensitive to extreme values is the *median.* In our example of ballplayers, the median is 189 pounds and it is the center of the continuum of the *ordered* weights of the players. In other words, half of the players have weights that are greater than 189 pounds and half of the players have weights that are less than 189 pounds. If the number of observations is an even number, the median is calculated by taking the simple arithmetic mean of the two middle values of the ordered observations. An extreme value for weight would not cause the median to shift along the continuum of ordered weights.

A third measure of central tendency is the *mode.* The mode is the most frequently occurring value in the data set. "By definition, the mode must be attained by more than one observation in the sample."[6] In our example, the mode is 200 pounds. That is, 200 pounds is the most frequently observed weight in the sample. The mode is not affected by extreme values in the data.

Of the three measures of central tendency, which is the most useful? Each tells us something about the weights of the ballplayers. For skewed samples, as a general guideline, the median is a better descriptive measure than the mean. However, many statistical tests such as the Student's t-test and ANOVA, assume an underlying Normal distribution of observations and this implied symmetry argues for the use of the mean.[2]

MEASURES OF SPREAD

Measures of spread reflect the *variability* in the data. Consider the following example modified from Remington

and Schork.[6] Suppose a marathon runner has just completed a race. Her feet are aching so she wishes to soak her feet after her ordeal. She locates two buckets of water each with a soothing, warm temperature of 78°F (26°C). Now consider the case in which she procures two buckets of water but one contains water at 34°F (1°C) for her left foot, and the second bucket contains water at 122°F (50°C) for her right foot. The result is that her left foot is practically freezing and her right foot is practically boiling. However, she should be "comforted" by the fact that "on average" she has an appropriate foot bath with a mean temperature of 78°F; the median is also 78°F. In fact, in both the first case and the second case the means and medians are identical (78°F), but obviously the situations are very different. The first situation has no variability in the temperature and the second situation has a great deal of temperature variation between the two buckets of water.

Specific measures of variation and spread affect our interpretation of the data and are important descriptors as well as being integral components of statistical testing. Measures of variation include the *range,* the *variance,* the *standard deviation,* and the *standard error.*

Refer to our example in Table 2 of the weights of ballplayers. The lowest weight in the data set is 140 pounds and the highest weight is 230 pounds. Frequently, investigators incorrectly report the range as two numbers from lowest value to the highest value of the observations in the data set. This is incorrect statistical usage because measures of variability are expressed as *single* numbers. The *range* is the highest weight in our sample minus the lowest weight, resulting in a difference, a range of 90 pounds. Likewise, with our marathon runner and the buckets of water, in the first case the range is 0° (78°F minus 78°F) and in the second case the range is 88°F (122°F minus 34°F). The range is the simplest measure of variation or spread. It tends to increase as the number of observations in a sample increase, and the range uses only the two extreme values of the data set and ignores ". . . all the information regarding variation that can be obtained from the remaining observations."[2]

A more informative measure of the variability of a sample is the *variance.* The variance is calculated by taking the difference between the mean value for a sample and the value for each particular individual in that sample, and then squaring and summing these differences. For example, consider again the weights of a group of 25 ballplayers, as summarized in Table 2. As already discussed, the range of the players' weight is 90 pounds, with a mean weight of 185.6 pounds, a median weight of 189 pounds, and a mode of 200 pounds. As can be seen from these data, some players deviate quite a bit from the mean. To determine the *variance* of this sample, each individual ballplayer's weight is subtracted from the mean weight of 185.6 pounds; the resultant difference from the mean value is squared, and these squared deviations added together; this sum (the summed deviations) is divided by the sample size minus one or "n−1." The resulting number, 381.32, is the sample variance.

The variance incorporates information for all of the weight observations in our sample. Unfortunately, the variance is not in the same physical units (pounds) as the weights in the sample of ballplayers; it is measured in *squared units.* In this example, 381.32 is *squared* pounds. To obtain a variability measure that is more interpretable, we use the square root of 381.32 square pounds, i.e., 19.52 pounds.

This square root of the variance is the *standard deviation* and is the most commonly reported measure of variability seen in the literature. We can tell a great deal about the variability in the sample by the size of the standard deviation. In our football player example, the standard deviation of 19.52 pounds is indicative of quite a bit of variation among the players in this sample.

Another measure of variability that is important to statistical inference is the *standard error.* The standard error is a measure of *the variability* from statistic to statistic, usually from mean to mean, i.e., variation from sample to sample. The standard error of the mean is the standard deviation divided by the square root of n. The standard error is used in calculations for such statistical tests as the Student's t-test, ANOVA, confidence intervals, and regression.

Information about the distribution of a variable, its central tendency, and its spread, taken together, tell us a great deal about a study. All of these measures are essential to our understanding of statistical analyses of continuous data.

ACKNOWLEDGMENT

The authors thank Dr. M. A. Schork from the University of Michigan, School of Public Health, Department of Biostatistics, for his review of this manuscript.

REFERENCES

1. Campbell MJ, Machin D: *Medical Statistics.* Chichester, England, John Wiley & Sons, 1990, p 393
2. Colton T: *Statistics in Medicine.* Boston, Little, Brown and Company, 1974, pp 30–31
3. Greenfield ML, Kuhn JE, Wojtys EM: A statistics primer. *Am J Sports Med 24:* 393–395, 1996
4. Hirsch RP, Riegleman RK: *Statistical First Aid: Interpretation of Health Care Data.* Boston, Blackwell Scientific Publications, 1992, p 22
5. Kuhn JE, Greenfield ML, Wojtys EM: A statistics primer: Statistical tests for discrete data. *Am J Sports Med 25:* 585–586, 1997
6. Remington RD, Schork MA: *Statistics with Applications to the Biological and Health Sciences.* Second edition. Englewood Cliffs, NJ, Prentice-Hall, Inc., 1985, p 15