Biostatistics and Microbiology

Biostatistics and Microbiology: A Survival Manual

Daryl S. Paulson
BioScience Laboratories, Inc.
Bozeman, MT, USA

Springer

Author
Daryl S. Paulson
BioScience Laboratories, Inc.
Bozeman, MT 59715
USA

ISBN: 978-0-387-77281-3 e-ISBN: 978-0-387-77282-0
DOI: 10.1007/978-0-387-77282-0

Library of Congress Control Number: 2008931293

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Acknowledgements

Researchers do not need to be statisticians to perform quality research, but they do need to understand the basic principles of statistical analysis. This book represents the most useful approaches learned through my years of experience in designing the statistically based studies in microbial death-rate kinetics, skin biotechnology, and clinical trials performed by my company, BioScience Laboratories, Inc.

I am truly grateful to my co-workers at BioScience, who kept the business running while I wrote and rewrote the chapters in this book, in particular, my very capable and supportive wife, Marsha Paulson, Vice-President of BioScience Laboratories, Inc. I am especially indebted to John Mitchell, Director of Quality Assurance, for his many hours of challenging my assumptions, utilizing his vast knowledge of microbiology, and editing this work. Tammy Anderson provided valuable assistance in the development of this book by typing and retyping it, making format changes, editing, creating figures and diagrams, and basically, managing the entire book development process. Her abilities are astounding.

Finally, I thank the staff at Springer for their patience, flexibility, professionalism, and quality concerns.

Contents

Chapter 1

BioStatistics and Microbiology: Introduction

To compete with the many books claiming to demystify statistics, to make statistics easily accessible to the "terrified," or provide an eastern approach purporting to present statistics that do not require computation, as in the "Tao of Statistics," is tough duty, if not utter fantasy. This book does not promise the impossible, but it will enable the reader to access and apply statistical methods that generally frustrate and intimidate the uninitiated. Statistics, like chemistry, microbiology, woodworking, or sewing, requires that the individual put some time into learning the concepts and methods. This book will present in a step-by-step manner, eliminating the greatest obstacle to the learner (not the math, by the way) applying the many processes that comprise a statistical method. Who would not be frustrated, when not only must the statistical computation be made, but an assortment of other factors, such as the α, β, and δ levels, as well as the test hypothesis, must be determined? Just reading this far, you may feel intimidated. I will counter this fear by describing early in the book a step-by-step procedure to perform a statistical method – a process that we will term "the six-step procedure." All of the testing will be performed adhering to six well-defined steps, which will greatly simplify the statistical process. Each step in the sequence must be completed before moving on to the next step.

Another problem that microbiology and other science professionals often must confront is that most of the training that they have received is "exact." That is, calculating the circumference of a circle tacitly assumes that it is a perfect circle; the weight of a material is measured very precisely to n number of digits; and 50 mL is, all too often, expressed to mean 50.0 mL exactly. This perspective of exactitude usually is maintained when microbiologists employ statistics; however, statistical conclusions deal with long-run probabilities which, by themselves, are nearly meaningless. In the context of microbiology, statistics can be extremely useful in making interpretations and decisions concerning collected data. Statistics, then, is a way of formally communicating the interpretation of clinical or experimental data and is particularly important when a treatment result is not clearly differentiable from another treatment. Yet, and this is the big "yet," the statistic used has much

D.S. Paulson, *Biostatistics and Microbiology*, doi: 10.1007/978-0-387-77282-0_1,

influence on the conclusions that result. A very "conservative" statistic requires very strong proof to demonstrate significant differences, whereas a "liberal" one requires less. "Yuck," you say already, "I just want to know the answer."

To this, I respond, when in doubt, use a conventional statistical method, one that can be agreed on and accepted by most authorities. These conventional kinds of methods will be presented in this book. As you gain experience, choosing statistical methods will become almost an intuitive process. For example, for problems in which you have little experience, you will be very cautious and conservative. By analogy, this is similar to rafting a river for the first time. If you see rapids in the river, you will be more conservative as you approach them – wearing a life jacket and helmet, and using your paddle to avoid rocks – at least until you have experienced them and developed confidence. You will tend to be more liberal when near a sandy shore in clear, calm, shallow water. For experiments in microbiology in which you have much experience, your microbiological knowledge enables you to be more statistically liberal, as you will know whether the result of statistical analysis is microbiologically rational.

Finally, statistics is not an end-all to finding answers. It is an aid in research, quality control, or diagnostic processes to support critical thinking and decision-making. If the statistical results are at odds with your field knowledge, more than likely, the statistical method and/or the data processed are faulty in some way.

1.1 Normal Distribution

Let's get right down to the business of discussing the fundamentals of statistics, starting with the normal distribution, the most common distribution of data. The normal distribution of data is symmetric around the mean, or average value, and has the shape of a bell. For example, in representing humans' intelligent quotients (IQs), the most common, or prevalent IQ value is 100, which is the average. A collection of many individual IQ scores will resemble a bell-shaped curve with the value 100 in the middle (Fig. 1.1).

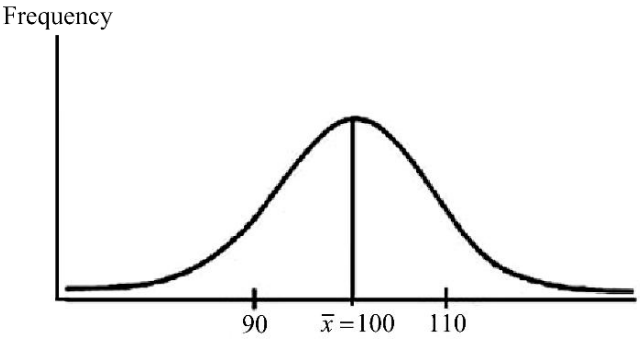

Fig. 1.1 Bell-shaped curve of intelligence quotients

IQ scores that are higher or lower than the mean are symmetrical around it. That is, values ten points above (110) and ten points below (90) the mean are equal distance from the mean, and this symmetrical relationship holds true over the entirety of the distribution. Notice also that, as IQ scores move farther from the mean in either direction, their frequency of occurrence becomes less and less. There are approximately the same number of 90 and 110 IQ scores, but far fewer 60 and 140 IQ scores.

Two values are important in explaining the normal distribution: the mean, or central value, and the standard deviation. When referring to an entire population, these values are referred to as *parameters*. When referring to a sample, they are referred to as *statistics*. The mean (average value) of a population is represented as μ, and the population standard deviation, as σ. For the most part, the values of the population parameters, μ and σ, are unknown for a total "population." For example, the true mean (μ) and standard deviation (σ) of the numbers of *Staphylococcus aureus* carried in the nasal cavities of humans are unknown, because one cannot readily assess this population among all humans. Hence, the statistical parameters, μ and σ, are estimated by sampling randomly from the target population. The sample mean ($\bar{x}$) and the sample standard deviation (s) represent unbiased estimates of the population parameters, μ and σ, respectively.

The mean $\bar{x}$ is the arithmetic average of values sampled from a population.* The standard deviation, s, describes how closely the individual values cluster around the mean value. For example, if all the measured values are within ± 0.1 g from the mean value in Test A and are within ±20.0 g from the mean value in Test B, the variability, or the scatter, of the data points in Test A is less than in Test B. The standard deviation is the value that portrays that range scatter, and does so in a very concise way. It just so happens that, in a large, normally-distributed data set, about 68% of the data are contained within ± one standard deviation (s) about the mean (Fig. 1.2).

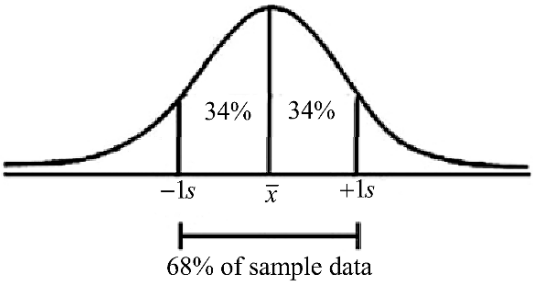

Fig. 1.2 Normally distributed data, ± one standard deviation from the mean

So, for example, if the mean number ($\bar{x}$) of *Staphylococcus aureus* colonies on 100 tryptic soy agar plates is 100, and the standard deviation (s) is 20, then 68% of the plate counts are between 80 and 120 colonies per plate (Fig. 1.3).

* Also, note that, for a theoretical normal distribution, the mean will equal the median and the mode values. The median is the central value, and the mode is the most frequently occurring value.

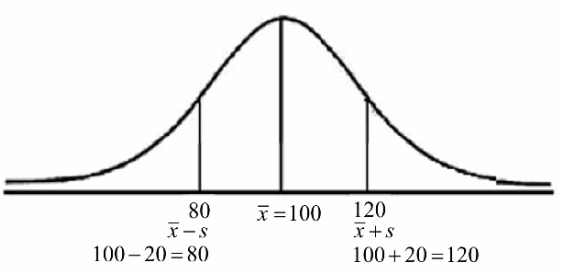

Fig. 1.3 Standard deviation of plate count values

If a second microbiologist counted colonies on the same 100 plates and also had an average plate count of $\bar{x}$ = 100, but a standard deviation (s) of 10, then 68% of the count values would be between 90 and 110 (Fig. 1.4).

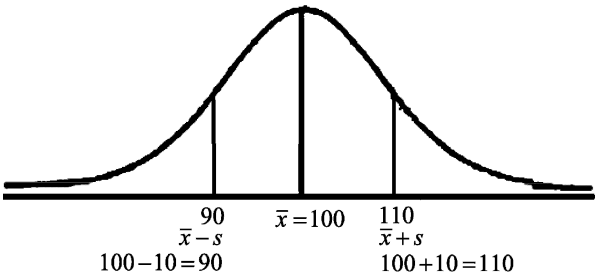

Fig. 1.4 Standard deviation of a second set of plate count values

The second microbiologist perhaps is more precise than the first in that the standard deviation, or scatter range of the data around the median, is smaller. On the other hand, he may consistently overcount/undercount. The only way to know is for both to count conjointly. Let's carry the discussion of standard deviations further.

± 1 standard deviation includes 68% of the data
± 2 standard deviations include 95% of the data
± 3 standard deviations include 99.7% of the data
Figure 1.5 provides a graphical representation.

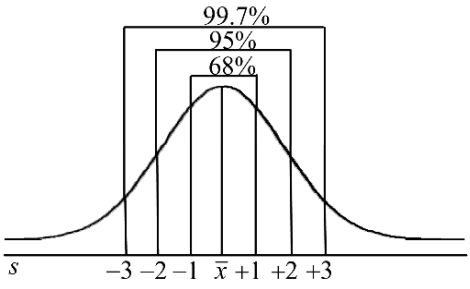

Fig. 1.5 Percentages of the area under the normal distribution covered by standard deviations

So, if one reads in a journal article that the $\bar{x}$ = 17.5 with an s of 2.3, one would know that 68% of the data lie roughly between 17.5 ± 2.3, or data points 15.2 to 19.8, and 95% lie between 17.5 ±2(2.3), or 12.9 to 22.1. This gives a person a pretty good idea of how dispersed the data are about the mean. We know a mean of 15 with a standard deviation of 2 indicates the data are much more condensed around the mean than are those for a data set with a mean of 15 and a standard deviation of 10. This comparison is portrayed graphically in Fig. 1.6.

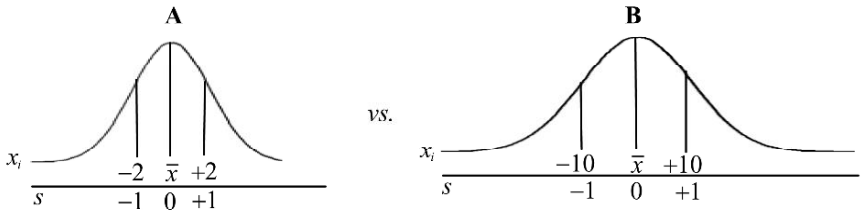

Fig. 1.6 a $\bar{x} = 15$ and $s = 2$, vs. **b** $\bar{x} = 15$ and $s = 10$

There is much more variability in the B group than in the A group. But then, what is variability? It is the scatter around the mean of the specific values.

Variability is measured by subtracting the mean of the data from each data point, i.e., $v = x_i - \bar{x}$. Take the data set [5, 6, 3, 5, 7, 4]. The mean is (5+6+3+5+7+4)/6 = 30/6 = 5. The variability of the data around the mean is $x_i - \bar{x}$.

$x_i - \bar{x}$	=	v
5 – 5	=	0
6 – 5	=	1
3 – 5	=	–2
5 – 5	=	0
7 – 5	=	2
4 – 5	=	–1
Sum	=	0

The variability is merely a measure depicting how far a data point is from the mean. Unfortunately, if one adds the individual variability points, v, they sum to 0. This makes sense in that, if the data are distributed symmetrically about the mean, there should be the same value weights more than and less than the mean that cancel each other out. A correction factor will be introduced in the next section so that the variability points around the mean will not cancel each other out. Variability is often interchanged with the term *statistical error*. Statistical error does not mean a mistake, or that something is wrong; it means that a data point differs from the mean.

There are times when statistical variability can mean missing a target value. For example, suppose I cut three boards 36 inches long. Here, the variability of the board lengths is the difference between the actual and target value. Suppose the boards actually measure 35.25 in., 37.11 in., and 36.20 in..

$$\begin{array}{lcll} 35.25 & - & 36.00 & = & -0.75 \\ 37.11 & - & 36.00 & = & 1.11 \\ 36.20 & - & 36.00 & = & 0.20 \end{array}$$

I guess this is why I am not a carpenter. The biggest difference between the variability of data around a mean and variability of data around a target value is that the sum of the individual points of difference, $x - \bar{x}$, using the mean will add to 0, but using the *x*-target value usually will differ from 0.

$$x_i - \bar{x} = -0.93 + 0.92 + 0.01 = 0$$
$$x_i - \text{target} = -0.75 + 1.11 + 0.20 = 0.56$$

Let us now discuss the concepts of mean and standard deviation more formally.

1.2 Mean

The mean, or arithmetic average, plays a crucial role. We are interested in two classes of mean value – a population mean and a sample mean.

μ = [mu] is the population mean. That is, it is the average of all the individual elements in an entire population – for example, the population mean number of bacteria found in the lakes of Wisconsin, or the population mean age of all the microbiologists in the world. Obviously, it is difficult, if not impossible to know the true population mean, so it is estimated by the sample mean, $\bar{x}$.

$\bar{x}$ = [*x* bar, or overline *x*] is the sample mean, or the arithmetic average of a sample that represents the entire population. Given the data are normally distributed, the sample mean is taken to be the best point estimator of the population mean. The calculation of the sample mean is $\bar{x} = (x_1 + x_2 + \ldots + x_n)/n$, where x_1 = the first sample value, x_2 = the second sample value, and $\ldots\ x_n$ = the last sample value. The subscript designates the sample number. More technically, $\bar{x} = \sum_{i=1}^{n} x_i \Big/ n$

Σ = [sigma] means "summation of." So anytime you see a Σ, it means add. The summation sign generally has a subscript, $i = 1$, and a superscript, n. That is, $\sum_{i=1}^{n} x_i$, where i, referring to the x_i, means "begin at $i = 1$" (x_1), and n means end at x_n.

The sub- and superscripts can take on different meanings, as demonstrated in the following. For example, using the data set:

i	x_i
1	6
2	7
3	5
4	2
5	3

i = 1 through 5, and as the last i, 5 also represents n. $\sum_{i=1}^{n} x_i = 6+7+5+2+3$; however, if $i = 3$, then we sum the x_i values from x_3 to $n = 5$. Hence,

$$\sum_{i=3}^{5} x_i = x_3 + x_4 + x_5 = 5+2+3$$

Likewise,

$$\sum_{i=1}^{n-1} x_i = x_1 + x_2 + x_3 + x_4 = 6+7+5+2$$

and

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 6+7+5$$

Often, for shorthand, we will use an unadorned sigma, $\sum$, to represent $\sum_{i=1}^{n}$ with any changes in that generality clearly signaled to the reader.

1.3 Variance and Standard Deviation

As stated earlier, the variance and standard deviation are statistics representing the difference of the x_i values from the $\bar{x}$ mean. For example, the mean of a data set, $\bar{x}$ = (118+111+126+137+148)/5 = 128.

The individual data point variability around the mean of 128 is

118 – 128 =	–10
111 – 128 =	–17
126 – 128 =	–2
137 – 128 =	9
148 – 128 =	20
Sum =	0

and the sum of the variability points $\sum(x_i - \bar{x})$ is zero. As previously noted, this makes sense – because the $\bar{x}$ is the central weighted value, this summation will never provide a value other than 0. So, we need to square each variability value $(x_i - \bar{x})^2$ and then add them to find their average. This average, $\frac{\sum(x_i - \bar{x})^2}{n}$, is referred to as the variance of the data.

$$\text{variance} = \frac{-10^2 + (-17)^2 + (-2)^2 + 9^2 + 20^2}{5} = \frac{874}{5} = 174.80$$

Likewise, the mean, the variance, and the standard deviation can be in terms of the entire population.

σ^2 = [sigma squared] = the true population variance $= \dfrac{\sum (x-\mu)^2}{N}$

where μ = true population mean, and N = true population size.

$\sigma = \sqrt{\sigma^2}$ = [sigma] = true population standard deviation

Again, because the entire population can rarely be known, the sample variance is computed as

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

Note that we divide by $n - 1$, not n. This is because we lose a degree of freedom when we estimate μ by $\bar{x}$.

$$s = \text{standard deviation} = \sqrt{s^2}$$

In hand-calculating s, a shortcut calculation is $s = \sqrt{\dfrac{\sum x^2 - n\bar{x}^2}{n-1}}$, which is generally faster to compute.

Two other important statistics of a data set are mode and median.

1.4 Mode of Sample Data

The mode is simply the most frequently-appearing numerical value in a set of data. In the set [4, 7, 9, 8, 8, 10], 8 is the mode.

1.5 Median of Sample Data

The median is the central numerical value in a set of data, essentially splitting the set in half. That is, there are as many individual values above it as below it. For data that are an odd number of values, it is the middle value of an ordered set of data. In the ordered data set [7, 8, 10], 8 is the median. For an ordered set of data even in the number of values, it is the sum of the two middle numbers, divided by 2. For the ordered data set [7, 8, 9, 10], the median is 8+9/2 = 8.5. This leads us directly into further discussion of normal distribution.

1.6 Using Normal Distribution Tables

As discussed earlier, the normal distribution is one of the most, if not the most, important distributions encountered in biostatistics. This is because it represents or models so many natural phenomena, such as height, weight, and IQs of individuals. Figure 1.7 portrays the normal distribution curve.

Because the normal distribution resembles a bell, it often is termed the "bell-shaped curve." The data have one central peak, that is, are "unimodal," and are symmetrical about the mean, median, and mode. Because of this symmetry, the mean = median = mode.

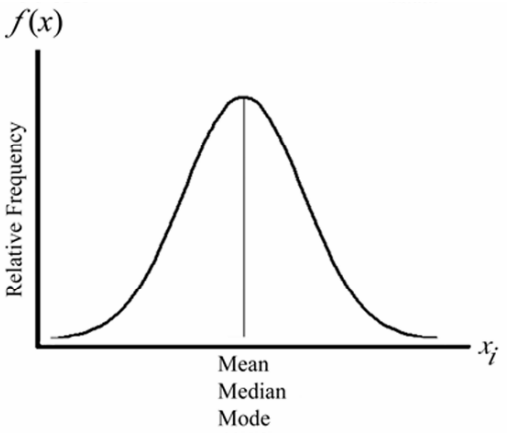

Fig. 1.7 A normal distribution, the "bell-shaped curve"

Most statistical methods associated with normal distributions utilize the mean and the standard deviation in their calculations. We already know that 68% of the data lie between $\mu + \sigma$ and $\mu - \sigma$ standard deviations from the mean, that 95% of the data lie between the mean and two standard deviations ($\mu + 2\sigma$ and $\mu - 2\sigma$, or $\mu \pm 2\sigma$), and 99.7% of the data lie within three standard deviations of the mean ($\mu \pm 3\sigma$). These relationships describe a theoretical population, but not necessarily a small sample set using the same mean ($\bar{x}$) and standard deviation (s). However, they are usually very good estimators. So, whenever we discuss, say, the mean ± 2 standard deviations, we are referring to the degree to which individual data points are scattered around the mean (Fig. 1.8).

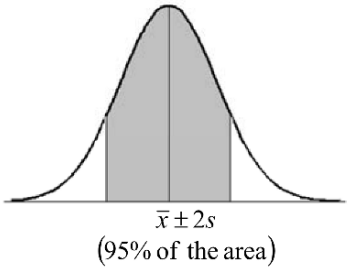

Fig. 1.8 The mean ± 2 standard deviations for a set of data

In any given sample, roughly ninety-five percent (95%) of the data points are contained within $\bar{x} \pm 2s$.

The z Distribution is used to standardize a set of data points so that the mean will be zero, and the standard deviation, 1, or $\bar{x} = 0$, $s = 1$. The transformation is $z = (x_i - \bar{x})/s$. A $z = 2.13$ means the value is 2.13 standard deviations to the right of the mean. A $z = -3.17$ means the value is 3.17 standard deviations to the left of the mean. The normal z Distribution table can be rather confusing to use, and we will use the Student's t Distribution in its place, when possible. This will not always be the case, but for now, we will focus on the Student's t Distribution. When the sample size is large, $n > 100$, the Student's t Table and the normal z Distribution are identical. The advantage of the Student's t Distribution is that it compensates for small sample sizes. The normal curve is based on an infinite population. Because most statistical applications involve small sample sizes, certainly fewer than infinity, the normal distribution table is not appropriate, for it underestimates the random error effect. The Student's t table (Table A.1) compensates for smaller samples by drawing the tails out farther and farther (Fig. 1.9).

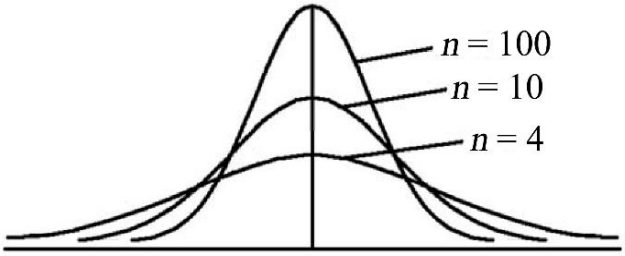

Fig. 1.9 Student's t Distribution versus the normal distribution

To use the Student's t Table (Table A.1), we need two values:

1. Sample size (n), and
2. α (alpha) level.

As you are now aware, the sample size is the number of experimental observations. The confidence level, $1 - \alpha$, is the amount of area under the distribution curve with which one is working. Generally, that value is 0.95; that is, it incorporates 95% of the area under the curve. The α level is the area outside the confidence area. If one uses two standard deviations, or a 95% confidence area, the α level is $1 - 0.95 = 0.05$. The $\alpha = 0.05$ means that 5% of the data are excluded. Note that there is nothing to figure out here. These are just statements.

In all the tests that we do using the Student's t table, the sample size will always be provided, as will the degrees of freedom. The degrees of freedom will be designated as df. If $df = n - 1$, and $n = 20$, then $df = 20 - 1 = 19$. Or, if $df = n_1 + n_2 - 2$, and $n_1 = 10$ and $n_2 = 12$, then $df = 10 + 12 - 2 = 20$. The smaller the df value, the greater the uncertainty, so the tails of the curve become stretched. In practice, this means the smaller the sample size, the more evidence one needs to detect differences between compared sets of data. We will discuss this in detail later.

The *df* value is used to find the Student's *t* tabled value at a corresponding α value. By convention, α is generally set at $\alpha = 0.05$. So, let us use Table A.1, when $\alpha = 0.05$ and for $n = 20$; hence, $df = n - 1 = 19$.

Using the *t* table:

Step 1. Go to Table A.1.

Step 2. Find *df* in the left column labeled *v*. Here, $v = 19$, so move down to $v = 19$.

Step 3. When you reach the 19, move right to the column corresponding with the α value of 0.05.

Step 4. Where the $v = 19$ row and the $\alpha = 0.05$ column meet, the tabled value = 1.729.

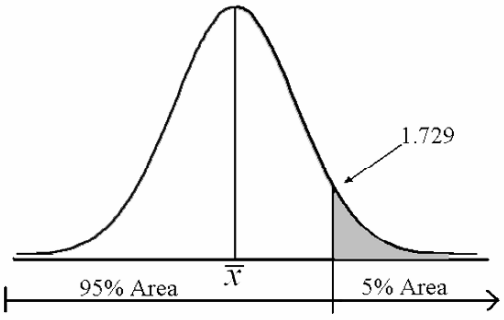

Fig. 1.10 The *t* tabled value = 1.729, an upper-tail value

The table, being symmetrical, provides only the upper (positive), right-side *t* tabled value. The 1.729 means that a *t* test value greater than 1.729 is outside the 95% confidence area (Fig. 1.10). This is said to be an upper-tail value. A lower-tail value is exactly the same, except with a minus sign (Fig. 1.11).

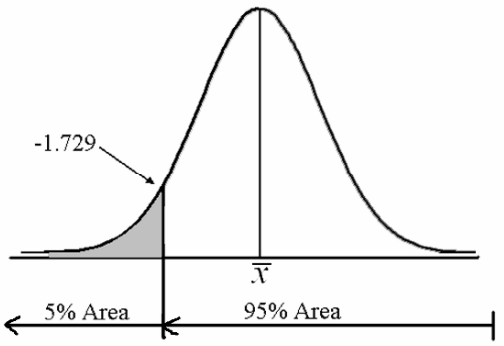

Fig. 1.11 The *t* tabled value = –1.729, a lower-tail value

This means that any value less than –1.729 is outside the 95% region of the curve and is significant. Do not worry about the *t* test values yet. We will bring everything together in the six-step procedure.

Finally, a two-tail test takes into account error on the upper and lower confidence levels. Here, you use two values of *a*, but you divide it, *a*/2. *a* = 0.05 = 0.05 2 = 0.25. Using the *t* table:

Step 1. Go to Table A.1.

Step 2. Find the *df* in the left column labeled v. Here, *v* = 19, so move down to *v*=19.

Step 3. when you read the 19, move right to the column corresponding to *a*/2 = 0.05 2 = 0.025.

Step 4. Where the *v* = 19 row and the *a*/2 = 0.025 column meet, the tabled value is 2.093. This means −0.2093 and 2.093.

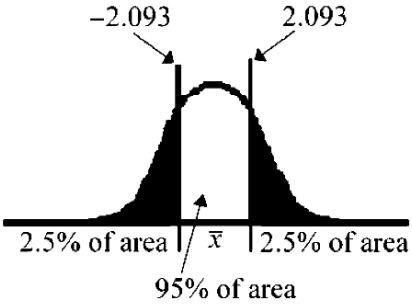

Fig. 1.12 Two-Tail Test

This table uses both values, −2.093 and 2.093. It is the upper and lower test jointly, with *a* being divided by 2, *a*/2 .

1.7 Standard Error of the Mean

To this point, we have discussed the variability of the individual x_i data around the mean, $\bar{x}$. Now, we will discuss the variability of the sample mean, $\bar{x}$ itself, as it relates to the theoretical ("true") population mean, μ. The standard deviation of the mean, *not the data points*, is also termed the standard error of the mean. In most statistical tests, the means of samples are compared and contrasted, not the data points themselves. Error, in this sense, is variability. The computation for the standard error of the mean, $s_{\bar{x}}$, is:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \text{ or } \sqrt{\frac{s^2}{n}}$$

What does the standard deviation of the mean represent? Suppose ten sample sets are drawn randomly from a large population. One will notice that the sample mean is different from one sample set to another. This variability is of the mean, itself. Usually, one does not sample multiple sets of data to determine the variability of the mean. This calculation is done using only a single sample, because a unique relationship exists between the standard deviation of the mean and the standard deviation of the data that is proportional to data scatter. To determine the standard error of the mean, the standard deviation value of the sample data is simply divided by the square root of the sample size, *n*.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The standard deviation of the mean has the same relation to the mean as the standard deviation of data has to the mean, except it represents the variability of the mean, not the data (Fig. 1.13). That is, for the mean $\bar{x} \pm s_{\bar{x}}$, 68% of the time the true mean, μ, is contained in that interval. About 95% of the time, the true mean value μ will lie within the interval, $\bar{x} \pm 2s_{\bar{x}}$. And about 99.7% of the time, the true mean value μ will lie within the interval, $\bar{x} \pm 3s_{\bar{x}}$.

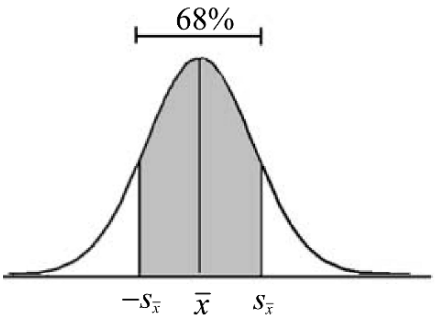

Fig. 1.13 The standard error of the mean

Key Points

s = Standard deviation of the data set	$s_{\bar{x}} = s/\sqrt{n}$ = Standard deviation, not of the data set, but of the mean

1.8 Confidence Intervals

Most of the work we will do with the normal distribution will focus on estimating the population mean, μ. Two common approaches to doing this are: 1) a point estimate and 2) calculation of a confidence interval estimate. The point estimate of μ is simply the sample mean, $\bar{x}$. The interval estimate of μ is $\bar{x} \pm t_{(\alpha/2,\, n-1)} s/\sqrt{n}$, with a confidence level of $1 - \alpha$.

Data Set. Suppose you are to estimate the true mean weight of each of 10,000 containers in a single lot of bacterial growth medium. Each container is supposed to contain 1 kg of medium. Ten are randomly sampled and weighed. The weights, in grams, are

n	x_i
1	998
2	1003
3	1007
4	992
5	985
6	1018
7	1009
8	987
9	1017
10	1001