

# THE UNIVERSITY OF TEXAS AT AUSTIN SCHOOL OF INFORMATION

## MATHEMATICAL NOTES FOR LIS 397.1 INTRODUCTION TO RESEARCH IN LIBRARY AND INFORMATION SCIENCE

Ronald E. Wyllys  
Last revised: 2003 Feb 12

### USING THE GAUSSIAN (NORMAL) DISTRIBUTION FOR APPROXIMATIONS

#### ESTIMATION VIA THE GAUSSIAN DISTRIBUTION

##### The Gaussian Distribution and the Distribution of Sample Means

The Gaussian, or normal, distribution works for a very large variety of statistical variables, in the sense that many such variables take on their values in patterns that are reasonably well approximated by the Gaussian distribution. Especially important among these variables is the sample mean, a variable that conforms very well to the Gaussian distribution.

One of the most frequently encountered situations in practical statistics is that in which we want to infer the location of the mean  $\mu_0$  of some population on the basis of the information obtained by observing a sample of the elements from the population. To draw such an inference involves reasoning based on the behavior of the set of *means of all samples of a given size,  $n$* , drawn from a population. Specifically, the reasoning employs the fact that the set of means of all samples of a given size behaves according to the Gaussian distribution, a fact that holds true even when the population itself is not distributed in a Gaussian fashion. This remarkable, but provable, fact--known as the Central Limit Theorem (or, more precisely, as the best known consequence of that theorem)--forms the basis for many of the everyday uses of statistics.

##### The Number of Distinct Samples

Closely linked to the fact that the Central Limit Theorem works is the following fact: The number of distinct samples of a given size (samples differing by at least one element) that can be drawn from a population is usually astronomical in size, even for relatively small samples and populations. To help you understand why this is so, here is an example.

Suppose that we had a population of 8 things, named, A, B, C, D, E, F, G, and H, respectively, and that we wanted to work out all the distinct samples of size 5 that could be chosen from this population. Bear in mind that what matters in distinguishing one sample from another is only what elements are in the samples, not the order in which the elements are drawn. The distinct samples are given below. Each individual sample is shown arranged in alphabetical order regardless of the sequence in which its elements were drawn, and the whole set is listed in an order (alphabetically down the columns) that makes it easy to keep track of what we are doing.

ABCDE	ABCFH	ABEFG	ACDFH	ADEGH	BCDGH	BDFGH
ABCDF	ABCGH	ABEFH	ACDGH	ADFGH	BCEFG	BEFGH
ABCDG	ABDEF	ABEGH	ACEFG	AEFGH	BCEFH	CDEFG
ABCDH	ABDEG	ABFGH	ACEFH	BCDEF	BCEGH	CDEFH
ABCEF	ABDEH	ACDEF	ACEGH	BCDEG	BCFGH	CDEGH
ABCEG	ABDFG	ACDEG	ACFGH	BCDEH	BDEFG	CDFGH
ABCEH	ABDFH	ACDEH	ADEFG	BCDFG	BDEFH	CEFGH
ABCFG	ABDGH	ACDFG	ADEFH	BCDFH	BDEGH	DEFGH

Note that these are distinct samples, because each of these samples differs by at least one element (i.e., by at least one letter) from every other sample. Altogether, there are 56 distinct samples of 5 elements each that can be drawn

from a population of 8 elements. In general, it can be shown that the number of distinct samples of size  $k$  that can be drawn from a population of  $N$  elements is given by

Equation 1. Number of ways of choosing  $k$  elements out of  $N$  elements

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

The symbol on the left side of this equation means the number of ways of choosing  $k$  things at a time out of a total of  $N$  different things; i.e., of choosing a sample of size  $k$  out of a set of size  $N$ . The symbol can be verbalized as "N choose  $k$ ." The "!" symbol on the right side is read as "factorial" (e.g., "N factorial"). The definition of  $N!$  is:

$$\begin{aligned} 0! &= 1 \\ 1! &= 1 \end{aligned}$$

and for  $N > 1$ ,

$$N! = N \times (N-1) \times (N-2) \times \dots \times 2 \times 1$$

That is,  $N!$  is  $N$  times each number less than  $N$  down to 1. For example,  $2! = 2$ ,  $3! = 6$ ,  $4! = 24$ ,  $5! = 120$ , ...,  $10! = 3,628,000$ .

You can see that as  $N$  increases,  $N!$  gets very large very fast indeed. This should suggest to you why the number of samples of size  $k$  that can be drawn from a population of size  $N$  is likely to be extremely large for typical values of  $k$  and  $N$  encountered in practical applications of inferential statistics. For the example above, the number of samples of size 5 that can be drawn from a population of 8 elements is

Equation 2. Number of samples of size 5 that can be drawn from 8 elements which matches the number of distinct samples shown above.

$$\binom{8}{5} = \frac{8!}{5!(8-5)!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1 \times (3 \times 2 \times 1)} = 8 \times 7 = 56$$

An example of the extremely large number of distinct samples possible with even moderate-size populations and samples is this: The number of distinct samples of size 50 that can be drawn from a population of size 1,000 is

$$9,460,461,017,585,217,846,063,722,277,728,044,918,729,694,001,668,654,064,793,- \\ 569,321,343,252,697,198,115,263,280$$

That is to say, this number is approximately 9,460 followed by 81 zeros or, in exponential notation,  $9.46 \times 10^{84}$ . This number is roughly the same size as the number of quarks in the entire known universe.

### Intervals for Elements of a Population

Now we return to the use of the Gaussian distribution in helping us determine the value of the mean of a population.

Suppose  $X$  is a variable for which the Gaussian distribution works. Usually, we shall be interested in determining a range of likely values for the (population) mean of  $X$ , either on the basis of theoretical considerations alone or by using information obtained from taking a sample and observing one or more values of  $X$  in the sample. In what follows "mean( $X$ )" and "SD( $X$ )" are to be read, respectively, as "mean of  $X$ " and "standard deviation of  $X$ ".

We can use the Gaussian distribution for the purposes just stated by noting that:

Equation 3. Area under the Gaussian curve

$$\Pr \left[ A < \frac{X - \text{mean}(X)}{\text{SD}(X)} < B \right] = \text{area under the Gaussian curve between A and B}$$

In general, the notation "P[something]" is to be read as "the probability of 'something'" or "the probability 'that something'." Thus Equation 3 can be read as implying the following: There is a certain probability that the quantity

$$\frac{X - \text{mean}(X)}{\text{SD}(X)}$$

has a value between numbers A and B (i.e., that this quantity is bigger than A but smaller than B). This probability is a number that is approximately equal to a particular area under the Gaussian curve (the "bell-shaped" curve), the area that is defined by having A as its left boundary and B as its right boundary.

The Gaussian curve in this formula is the "standardized" Gaussian curve, whose mean value is 0 and whose standard deviation is 1. This standardized Gaussian curve is the one for which all tables of the Gaussian curve are constructed. What the

$$\frac{X - \text{mean}(X)}{\text{SD}(X)}$$

term does is to standardize the variable  $X$  in a similar way. In detail, subtracting the mean of  $X$  from  $X$  gives us an intermediate variable,  $X - \text{mean}(X)$ , whose mean is zero; and dividing that intermediate variable by the standard deviation of  $X$  gives us still another variable

$$\frac{X - \text{mean}(X)}{\text{SD}(X)}$$

whose standard deviation is one (and whose mean remains zero).

If two qualifications are met, viz., if (i) we know that variable  $X$  is reasonably well approximated by the Gaussian distribution, and if (ii) we know both the mean and the standard deviation of  $X$ , then we can use formula (2.1) to give us a region of likely values for  $X$ . For example, suppose we want to find the region that will take in 95% of the values of  $X$ . We know that the area under the Gaussian curve between -2 and +2 is approximately .95, so we use  $A = -2$  and  $B = 2$  in Equation 3. That is, we can write:

$$\Pr \left[ -2 < \frac{X - \text{mean}(X)}{\text{SD}(X)} < 2 \right] = \text{area under the Gaussian curve between } -2 \text{ and } 2$$

from which it follows that we want to find those values of  $X$  such that:

$$-2 < \frac{X - \text{mean}(X)}{\text{SD}(X)} < 2$$

Equivalently, by multiplying throughout by  $\text{SD}(X)$ , we have

$$-2\text{SD}(X) < X - \text{mean}(X) < 2\text{SD}(X)$$

Equivalently again, by adding  $\text{mean}(X)$  throughout, we obtain

Formula 4. Boundary values for  $X$

$$\text{mean}(X) - 2\text{SD}(X) < X < \text{mean}(X) + 2\text{SD}(X)$$

We can say that Formula 4 tells us that we are interested in those values of  $X$  that are somewhere between the following two numbers:  $\text{mean}(X) - 2\text{SD}(X)$ , on the left, and  $\text{mean}(X) + 2\text{SD}(X)$ , on the right.

For a variable  $X$  that meets qualifications (i) and (ii), Formula 4 gives us what is called a "95%

confidence interval," i.e., an interval that can be expected to contain 95% of the values of  $X$ . To put it another way, if we made a random selection of a value of  $X$ , the probability that that value would lie within the interval would be 95%. Other confidence intervals for  $X$  would come from other values of A and B.

(Note: As the foregoing comments suggest, a better name for such intervals would be "probability interval." Some people argue strongly for using the latter name. However, the term "confidence interval" and such phrases as "we are 95% confident that X is in the interval..." are ineradicably fixed in the language, and we will use them in these *Mathematical Notes for LIS 397.1, Summer 1998.*)

Since we ordinarily are interested in intervals that are *symmetric* around  $\text{mean}(X)$ , we can ordinarily let  $A = -C$  and  $B = C$ , and use the limits:

Formula 5. Symmetric boundary values for  $X$

$$\text{mean}(X) - \text{CSD}(X) < X < \text{mean}(X) + \text{CSD}(X)$$

Formula 5 gives us the confidence interval whose probability (i.e., whose confidence coefficient) equals the area under the Gaussian curve from  $-C$  to  $C$ , when  $X$  satisfies qualifications (i) and (ii). This confidence interval is then the probability that  $X$  takes on values within  $C$  standard-deviation units of  $\text{mean}(X)$ , or to put it in still other words, that a particular observed value of the variable  $X$  is no farther than  $C$  standard deviations away from the mean of the variable,  $\text{mean}(X)$ .

Formula 5 may be re-written as

Formula 6. Statement of a symmetric interval containing  $X$

$$X \text{ is in the interval, } \text{mean}(X) \pm \text{CSD}(X)$$

Note that Formula 6 has the form:

Formula 7. Confidence interval statement of a symmetric interval containing  $X$

$$(\text{mean of something}) \pm [(\text{confidence factor}) \text{ times } (\text{SD of the something})]$$

As an illustration we can use an IQ score. Because of its design, the Wechsler Adult Intelligence Score (WAIS) has a mean of 100 and a standard deviation of 15. This means that if you took a large random sample of adults, you would expect to find that about 68% of them would have a  $\text{WAIS} = X$  in the range

$$100 - 1(15) < X < 100 + 1(15)$$

or, equivalently, for 68% of the  $X$ s (WAISs) it would be true that  $85 < X < 115$ .

Similarly, about 95% of the adults' WAISs would lie in the range:

$$100 - 2(15) < X < 100 + 2(15)$$

or, equivalently, for 95% of the  $X$ s (WAISs) it would be true that  $70 < X < 130$ .

Note that here we are using knowledge of the population to make inferences about the sample. This is just the opposite of what we ordinarily want to do in using statistical reasoning; but it is instructive to see what happens in reasoning from the population to a sample, as an aid in understanding how to reason from a sample to the population.

We can use the same illustration to go a step further. What we just did was to show that there is a probability of 0.68 that an adult chosen at random would have a WAIS greater than 85 but less than 115 (and similarly, a probability of 0.95 of a WAIS between 70 and 130). A different kind of question is this: If we take a sample of  $n$  adults, what can we say about a probable range of values for the *mean WAIS of the sample group*? Now we are interested in a new variable:  $\bar{X}$ , the sample mean.

## Intervals within which Sample Means Lie

The reason for our being concerned with a sample is ordinarily that we are in the process of investigating some population that is currently of interest to us. For convenience in the discussion that follows, we shall give this population the name, "Population COI," where "COI" comes from "currently of interest."

The sample mean is indeed a variable, for if we took many different samples, each of size  $n$ , from Population COI, we would get many different values for the means of the samples. Suppose we actually did this for, say,  $K$  different samples. We would get  $K$  observations,  $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_K$ , of values of the sample mean, and they would, in general, be different from one another. Now suppose that  $K$  is that particular number (in general, an enormous number) that represents the number of *all possible* samples of size  $n$  drawn from Population COI. It is important to realize that the set of  $K$  sample means is itself a population. For convenience, we shall call it "Population APS," where "APS" comes from "all possible samples."

With sufficient time and patience, we would be able to verify certain important relationships between Populations APS and COI. First, we would find that the mean of Population APS, i.e., the *mean of the set of means of all possible samples of size  $n$  drawn from Population COI*,

$$\bar{\bar{X}} = \left(\frac{1}{K}\right) \sum \bar{X}_i$$

would be equal to the mean,  $\mu_0$ , of Population COI.

Second, we could directly calculate the standard deviation of Population APS, the *standard deviation of the set of means of all possible samples of size  $n$  drawn from Population COI*, by using the formula for the standard deviation of a population. If we carried out this calculation (which, in general, would take an enormously long time even on a supercomputer), we would be in a position to verify that this population standard deviation can also be calculated by a far simpler method, which we explain below. The correctness of the simpler method can be proven theoretically (though we shall not do so here).

The standard deviation of Population APS is of such great importance in inferential statistics that it has been given a special name, "the standard error of the mean," which is often shortened to "standard error" or even just "SE." It has also been given a special notation,

$$\sigma_{\bar{X}}$$

or

$$\sigma_{\bar{X}(n)}$$

The simpler calculation mentioned above consists in the fact that the SE is equal to the standard deviation,  $\sigma$ , of Population COI, divided by the square root of  $n$ , the size of the samples making up Population APS:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Part of this relationship should seem intuitively reasonable: The sample means are averages, and the averaging process produces results that tend to be in the middle of whatever range of values is possible for the individual observations. That is, the sample means should cluster around their own mean, which happens to be the mean of Population COI. Hence, the average spread of the *sample means* (which is what the standard error of the mean,  $\sigma_{\bar{X}}$ , represents) should be smaller than the average spread of the individual *elements* of Population COI (which is what the standard deviation of this population,  $\sigma$ , represents).

What may not seem intuitively reasonable, and is indeed remarkable, is that the shape of the clustering is Gaussian. That is, if you measured the various values of the means of many samples of size  $n$  and constructed a histogram for

the values of these means, you would find that the histogram looked like a bell-shaped curve centered on the population mean.

Strictly speaking, the relationship between  $\sigma_{\bar{X}}$  and  $\sigma / \sqrt{n}$  is that of approximate equality. However, the approximation is very good provided that  $n$  (the size of the sample) is such that  $n > 30$  and that  $n$  is less than 5% of the size of the population. These conditions are easily met in practice, and hence for convenience we shall ordinarily treat the relationship as an exact equality.

Note that  $\sigma_{\bar{X}}$  represents the *true* standard error, i.e., the SE of Population APS, and that  $\sigma$  represents the *true* standard deviation of Population COI. In real life, we almost never know exactly what  $\sigma$  is, because we lack complete information about Population COI. (After all, if we already knew everything about this population, we would not have to be using the techniques of inferential statistics in order to gain information about it.)

In practice, what we do is use the observed sample standard deviation,  $s$ , as the best available approximation to  $\sigma$ . When we use  $s$  in place of  $\sigma$  in the equation for the SE, we have the relation

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

and it is this relation that provides the numerical value for the standard error in almost all practical cases.

Now we are ready to apply Equation 3. The variable  $X$  of the equation becomes our variable  $\bar{X}$ , the sample mean (remember that we are dealing here with the means of all possible samples of size  $n$ );  $\text{mean}(X)$  becomes  $\mu$ , the population mean; and  $\text{SD}(X)$  becomes  $\sigma_{\bar{X}}$ , the standard error of the mean. We have:

$$\begin{aligned} \text{area under the Gaussian curve} \\ \text{between A and B} \end{aligned} = \Pr \left[ A < \frac{X - \text{mean}(X)}{\text{SD}(X)} < B \right] = \Pr \left[ A < \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < B \right]$$

and, since  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$  we have

$$\text{area under the Gaussian curve between A and B} = \Pr \left[ A < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < B \right]$$

By the same kind of reasoning that led to Formulas 4 and 5 we can say that the following limits for  $\bar{X}$

$$\mu - A \left( \frac{\sigma}{\sqrt{n}} \right) < \bar{X} < \mu + B \left( \frac{\sigma}{\sqrt{n}} \right)$$

give us that particular confidence interval, for the sample mean, whose confidence coefficient equals the area under the Gaussian curve from A to B. As before, we are ordinarily interested in that are symmetric around the mean, so that we can ordinarily let A = -C and B = C, and write

Formula 6a. Symmetric limits for the sample mean

$$\mu - C\left(\frac{\sigma}{\sqrt{n}}\right) < \bar{X} < \mu + C\left(\frac{\sigma}{\sqrt{n}}\right)$$

Equivalently, we have

Formula 6b. Symmetric limits for the sample mean

$$\bar{X} \text{ is in the interval, } \mu \pm C\left(\frac{\sigma}{\sqrt{n}}\right)$$

These formulas specify the confidence interval (more precisely, the symmetric confidence interval) for the sample mean whose confidence coefficient equals the area under the Gaussian curve from -C to +C. Probably the most frequently used value of C is 1.96; the area under the Gaussian curve from -1.96 to 1.96 is .95; and thus the 95% confidence interval is the one calculated by using the value C = 1.96 in Formulas 6a and 6b.

Let us look at an example. If we took a large number of samples, each consisting of 25 adults, and if for each sample we found the WAISs of the adults in the sample and then calculated the mean WAIS for the sample, we would expect to find that about 68% of the sample means would lie in an interval calculated by using C = 1 in Formula 6a:

$$\mu - 1\left(\frac{\sigma}{\sqrt{n}}\right) < \bar{X} < \mu + 1\left(\frac{\sigma}{\sqrt{n}}\right)$$

Using the actual numbers, we have

$$100 - 1\left(\frac{15}{\sqrt{25}}\right) < \bar{X} < 100 + 1\left(\frac{15}{\sqrt{25}}\right)$$

$$100 - 3 < \bar{X} < 100 + 3$$

$$97 < \bar{X} < 103$$

Using Formula 6b we could also state the result this way: For 68% of the sample means,  $\bar{X}$ , it is true that

$$\bar{X} \text{ is in the interval, } \mu \pm 1\left(\frac{\sigma}{\sqrt{n}}\right)$$

or, using the actual numbers,

$$\bar{X} \text{ is in the interval, } 100 \pm 1 \left( \frac{15}{\sqrt{25}} \right)$$

$$\bar{X} \text{ is in the interval, } 100 \pm 3$$

### Confidence Intervals for Population Means

Note that the result in the preceding example does *not* say that 68% of the adults in a sample will have WAISs between 97 and 103. (Actually, of course, the range for 68% of the adults would be 85 to 115, as we saw earlier.) It says, instead, that the *mean* of the WAISs of a randomly chosen sample of 25 adults will, with probability 0.68, lie in the interval (97, 103). In general, the *mean of a randomly chosen sample of size n*, drawn from a population whose mean is  $\mu$  and whose standard deviation is  $\sigma$ , will, with probability 0.68, lie in the interval between

$$\mu - 1 \left( \frac{\sigma}{\sqrt{n}} \right) \text{ and } \mu + 1 \left( \frac{\sigma}{\sqrt{n}} \right)$$

that is, in the interval,  $\mu \pm 1 \left( \sigma / \sqrt{n} \right)$ , as stated in Formulas 6a and 6b. Other probabilities would correspond to other values than 1 for the multiplier of the  $\sigma / \sqrt{n}$  term.

What we have just been talking about is the expected placement of the mean of a sample drawn from a population whose parameters are *known*. Our information about this situation can be turned around to help us in the opposite situation.

The opposite situation, which is the usual one in practice, is that we do *not* know the population parameters. As we have just seen, in the *unusual* situation of *known* parameters we make statements like:

Approximately 95% of *all possible* sample means will lie centered around the population mean, within a distance, from the population mean, of 2 standard deviations of the set of means of all possible samples.

(We could omit the word "approximately" if we used the exact value 1.96 for the number of standard deviations.) In the *usual* situation of *unknown* parameters we make statements like:

We assume that the *observed* sample mean is one of the central 95% of all possible sample means. If so, it lies within a distance, from the unknown population mean, of 2 standard deviations of the set of means of all possible samples, i.e., within a distance of 2 standard errors (2 SEs) of this unknown population mean. That is, we assume that the population mean is at a distance of at most 2 SEs away from the observed sample mean. In other words, we assume that the population mean lies within an interval extending from 2 SEs below the observed sample mean to 2 SEs above the observed sample mean.

By this change in our point of view, we move from Formulas 6a and 6b to analogous formulas:

Formula 7a. Boundaries for the population mean in terms of observed sample mean and known population standard deviation

$$\bar{X} - C \left( \frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{X} + C \left( \frac{\sigma}{\sqrt{n}} \right)$$

Formula 7b. Interval for the population mean in terms of observed sample mean and known population standard deviation

$$u \text{ is in the interval, } \bar{X} \pm C \left( \frac{\sigma}{\sqrt{n}} \right)$$

We use Formulas 7a and 7b if we *know* the population standard deviation,  $\sigma$ . But ordinarily we *do not know*  $\sigma$ . Therefore, we ordinarily must settle for the best estimate of it, which is the sample standard deviation,  $s$ . Thus ordinarily we use one of the following formulas:

Formula 8a. Boundaries for the population mean in terms of observed sample values

$$\bar{X} - C \left( \frac{s}{\sqrt{n}} \right) < \mu < \bar{X} + C \left( \frac{s}{\sqrt{n}} \right)$$

Formula 8b. Interval for the population mean in terms of observed sample values

$$u \text{ is in the interval, } \bar{X} \pm C \left( \frac{s}{\sqrt{n}} \right)$$

In Formulas 7a and 7b the  $\sigma / \sqrt{n}$  factor represents the *true* standard error of the mean, i.e., the standard error of the mean calculated theoretically on the basis of (assumed) knowledge of the population. This standard error is what we use  $\sigma_{\bar{X}}$  to represent.

Analogously, in Formulas 8a and 8b, the  $s / \sqrt{n}$  factor represents the *best available estimate* of  $\sigma_{\bar{X}}$  on the basis of the information available in the sample you have in hand. This estimate, which is often denoted by  $s_{\bar{X}}$ , is what is ordinarily referred to as the *standard error of the mean*. Thus, in other discussions of this topic, you will encounter such formulas as

Formula 9a.

$$\bar{X} - C\sigma_{\bar{X}} < \mu < \bar{X} + C\sigma_{\bar{X}}$$

This can be re-stated as

Formula 9b.

$$\mu \text{ is in the interval, } \bar{X} \pm C\sigma_{\bar{X}}$$

We also have

Formula 10a.

$$\bar{X} - Cs_{\bar{X}} < \mu < \bar{X} + Cs_{\bar{X}}$$

Formula 10b.

$$\mu \text{ is in the interval, } \bar{X} \pm Cs_{\bar{X}}$$

The factors  $\sigma / \sqrt{n}$  and  $s / \sqrt{n}$  represent the "standard deviation of the something" part of Formula 7. The multiplier C is the confidence factor, a number that incorporates the level of confidence, as noted in Formula 7b. That is, C is a number chosen in such a way as to reflect the probability that the observed sample mean  $\bar{X}$  is one of the so-many-percent of all possible sample means that fall within the specified distance from the true population mean.

We can put Formula 10b into words as follows:

We construct an interval for the population mean by calculating

$$(\text{sample mean}) \pm [(\text{confidence factor}) \times (\text{standard error of the mean})]$$

Then our confidence that the population mean is in that interval is the numerical percentage that corresponds to the confidence factor we use.

You should also compare Formulas 10a and 10b with the discussion on pages 190-192 of the Mendenhall text.

### Examples of Estimation

These ideas may become clearer if we pause for numerical examples.

#### Example 1. An Example with Known Population Standard Deviation

For convenience, we confine ourselves initially to the situation in which the population standard deviation,  $\sigma$ , is *known*. Thus Formulas 7 and 9 will apply.

First, we need to find a variable for which it is realistic to assume that we could know the variable's standard deviation while failing to know its mean. Such an example is furnished by the WAIS. It happens that the mean WAISs of different subsets of the population of adults tend to differ much more than do the standard deviations of such subsets. That is, if one looks at a subset of adults, such as college graduates, one expects to find that their *mean* WAIS will be *different* from the *population mean*, 100, but that the *spread* of the WAISs of the college graduates around their mean WAIS will be about the *same* as the *spread* of WAISs in the population of all adults. In short, we can assume that the population standard deviation,  $\sigma$ , of the WAISs is known for any ordinary subset of adults.

In particular, we can assume that the standard deviation of the WAISs of GSLIS students is 15, which is the standard deviation of the WAIS because of the way the WAIS is constructed. Taking this group of students to be the population of interest, we could choose a random sample of, say, 16 GSLIS students, administer a Wechsler test, and determine the mean WAIS of the sample.

Suppose that we have found the mean WAIS of the sample to be 120, and that on the basis of this finding we would like to be able to make an inference about the probable value of the mean WAIS of the GSLIS student body. We are in the position of having observed a particular one of the many possible means of samples of size 16 that could be drawn from the population of GSLIS students. That is, the variable we have observed is the sample mean, and we know that the standard error of the mean is  $\sigma_{\bar{X}} = \sigma / \sqrt{n}$ . We have  $\sigma_{\bar{X}} = 15 / \sqrt{16} = 15 / 4 = 3.75$ . Thus using Formula 9a we can say that the 95% confidence limits for the population mean are:

$$\bar{X} - 1.96\sigma_{\bar{X}} < \mu < \bar{X} + 1.96\sigma_{\bar{X}}$$

$$120 - 1.96(3.75) < \mu < 120 + 1.96(3.75)$$

$$112.65 < \mu < 127.35$$

Alternatively, we could use Formula 9b and write that we have 95% confidence that

$$\mu \text{ is in } \bar{X} \pm 1.96\sigma_{\bar{X}} = 120 \pm 1.96(3.75)$$

Equivalently, we could say that  $\mu$  is in (112.65, 127.35) with 95% confidence.

Thus we have found the interval (112.65, 127.35) to be the 95% confidence interval for the mean WAIS of all GSLIS students. At any rate, it is the 95% confidence interval so far as we can tell from this particular sample.

### Example 2. An Example with Unknown Population Standard Deviation

Now let us turn to the usual situation in practice, the situation in which we do *not* know the population standard deviation,  $\sigma$ . In this situation we must rely on our best available estimate of  $\sigma$ , which is the observed standard deviation in the sample in hand,  $s$ . In this case, Formulas 8 and 10 apply, as we shall illustrate.

Suppose we have taken a random sample of 49 GSLIS women students (or, more precisely, a random sample of the heights of such women) and have found that they have a mean height of 66.34 inches with a standard deviation of 2.24 inches. To obtain the 95% confidence interval for the mean height,  $\mu$ , of the population of all GSLIS women students, we put these values into Formula 8a.

We find

$$66.34 - 1.96\left(\frac{2.24}{\sqrt{49}}\right) < \mu < 66.34 + 1.96\left(\frac{2.24}{\sqrt{49}}\right)$$

$$66.34 - 1.96\left(\frac{2.24}{7}\right) < \mu < 66.34 + 1.96\left(\frac{2.24}{7}\right)$$

$$66.34 - 1.96(0.32) < \mu < 66.34 + 1.96(0.32)$$

$$66.34 - 0.63 < \mu < 66.34 + 0.63$$

Thus we can say with 95% confidence, " $\mu$  is in (65.71, 66.97)"; or, slightly more precisely, we can say, "The probability is .95 that  $\mu$  is in (65.71, 66.97)."

### The Rôle of the t-Table

Only one more detail remains to be mentioned. It applies to the situation which, you are reminded once again, is by far the most frequent one in practice: This is the situation in which we do not know the population standard deviation,  $\sigma$ , and hence must work with the estimate of it provided by the information in the sample, viz., the sample standard deviation,  $s$ .

Naturally, there will be variation from sample to sample in the extent to which the observed values in a sample reflect the dispersion characteristic of the population. That is, different samples provide different values of the sample standard deviation,  $s$ , and hence, different estimates of the population standard deviation,  $\sigma$ . The larger the size of the samples, the less will be the variation in  $s$  among the samples, so that for "sufficiently large" samples, the variation can be ignored. (There are various rules-of-thumb for what constitutes "sufficiently large," but the threshold for a sufficiently large sample is usually considered to be somewhere in the range of 31 to 60.)

For samples whose size is not "sufficiently large," the sample standard deviation  $s$  in some of the samples will be enough smaller than  $\sigma$  to cause a problem. You can see from Formulas 8 that a too small value of  $s$  would yield a too small confidence interval for the population mean. To compensate, in working with small samples, one obtains the confidence factor from a table of the Student's  $t$  distribution instead of from a table of the Gaussian, or  $z$ , distribution.

(Note: The distribution known as "Student's  $t$ " was first studied by William S. Gossett, who published a paper on it in 1908 under the pseudonym, "Student." He used a pseudonym because his employer, the Guinness brewing company, did not want his work to be traced back to the company. Gossett was, in fact, doing what is now known as quality-control engineering, which we can all recognize as desirable in the brewing of beer. The company, no doubt rightly, felt that his work gave them a competitive edge; and they wanted to keep secret the fact that statistics can help brew better beer.)

In the table of the Student's  $t$  distribution in the Mendenhall text (p. 486), confidence factors appropriate for samples of size  $n$  are tabled in the row indicated by  $n-1$  df (degrees of freedom). The column headed  $t_{.025}$  contains the values corresponding to 95% confidence intervals, and the column headed  $t_{.005}$  contains the values corresponding to 99% confidence intervals

In the situation in which the population standard deviation,  $\sigma$ , is not known, the t-table *must* be used for samples whose size  $n$  is 30 or less, and it is *preferable* to use it, rather than the Gaussian table, for samples up to  $n = 100$  or more (depending on the t-table available).