

# 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 Jan 15

### CONFIDENCE INTERVALS FOR THE MEAN OF A POPULATION

#### INTRODUCTION

This MathNote deals with the construction of confidence intervals for the mean of a population. We assume that you are interested in acquiring information about some population, and that you have drawn a suitable (i.e., adequately randomized) sample from this population, a sample containing  $n$  observations.

What can you do with this in-hand sample of size  $n$  to get an idea of what the mean and the standard deviation of the population are?

The simplest answers are: First, the mean of your sample,  $\bar{x}$ , is the best piece of evidence you have as to what the population mean,  $\mu$ , is. That is,  $\bar{x}$  is your best **point estimate** of the population mean,  $\mu$ . Second, the standard deviation of your sample,  $s$ , is the best piece of evidence you have concerning the population standard deviation,  $\sigma$ . That is,  $s$  is your best point estimate of  $\sigma$ .

Point estimates are better than no information at all, but in practice we usually prefer to make use of a still better statement about the population mean than the point estimate. This MathNote discusses this preferred kind of statement, confidence intervals for the population mean. (It is worth noting that confidence intervals can also be constructed for the population standard deviation, but in practice this is rarely done.)

#### CENTRAL LIMIT THEOREM

To develop the notion of a confidence interval for a population mean, we begin by noting what the Central Limit Theorem (CLT) tells us. From the CLT we learn that sets of a certain type behave in an amazing way. The type we are talking about are sets that consist of the **means of all possible samples of a given size drawn from a particular population**. The CLT says that the means in any such set cluster in a Gaussian (i.e., bell-shaped curve) fashion around the mean of the population from which the samples are drawn. That is, (1) the mean of the whole set of sample means is equal to the population mean,  $\mu$ ; (2) most of the sample means lie close to  $\mu$ ; (3) as you move farther and farther away from  $\mu$  in either direction, below it or above it, you will find fewer and fewer sample means; and (4) this decrease occurs in a bell-shaped curve fashion as you move away from  $\mu$ .

#### STANDARD ERROR OF THE MEAN

Furthermore, it can be shown that the standard deviation of such a set of sample means is related to the standard deviation of the population,  $\sigma$ , in a surprisingly simple and elegant way. To denote the standard deviation of the set of means of samples of size  $n$  drawn from a particular population, we use the symbol

$\sigma_{\bar{x}(n)}$ ; and we give this special standard deviation a special name, the **standard error of the mean**, in recognition of its importance. (I recommend that you not try to parse the phrase, "standard error of the mean"; you should simply accept it as a label for a particular concept, just as you accept the label "Long-horns" for a certain group of football players, despite their hornlessness.) The elegant relationship is That is, to calculate the standard error of the mean, you simply divide the standard deviation of the

$$\sigma_{\bar{x}(n)} = \frac{\sigma}{\sqrt{n}}$$

population,  $\sigma$ , by the square root of  $n$ , the number of elements in the samples that gave rise to the set of means about which we have been talking.

## DETERMINING THE STANDARD DEVIATION OF THE POPULATION

A natural question is: Since we are presumably trying to study an unknown population in order to find out what it is like, how can we know what its standard deviation is?

The answer to this question is twofold: First, sometimes you know what the standard deviation of a population is even though you do not know what the mean of that population is. An example is a population that consists of scores on a standardized test, such as the Wechsler Adult Intelligence Scale (WAIS), which by design have been arranged to have a standard deviation of 15 and an overall mean of 100. For example, an investigator might be interested in comparing the mean WAIS scores of various subpopulations (of the population of all adults), such as library and information science students, on the one hand, and business-school students, on the other hand. It is a fact that, with respect to almost any measure, the means of different subpopulations can differ much more easily than do the standard deviations of different subpopulations. (This fact is often referred to as the "stability" of the standard deviation [and of the variance].) Because of the stability of the standard deviation, the investigator could safely proceed on the assumption that although LIS and business students might have different WAIS means, the two groups would almost surely both have a WAIS standard deviation of 15.

Second, in the more general case where you do not know the standard deviation of the population, the stability of the standard deviation makes it reasonable to accept the observed sample standard deviation as likely to be close enough, for most practical purposes, to the population standard deviation for you to go ahead and work with it. That is, you are usually justified in treating the observed sample standard deviation,  $s$ , as equal to  $\sigma$  for most practical purposes. In particular, for most practical purposes you can work with the relationship,

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

and trust it to give you useful results. In the relationship just displayed, we have defined a new symbol,  $s_{\bar{x}(n)}$ , which we can call the observed standard error of the mean.

## CALCULATING CONFIDENCE INTERVALS FOR THE POPULATION MEAN

Because the set of means of all possible samples (of a given size, drawn from a particular population) is distributed around the population mean,  $\mu$ , in a Gaussian fashion--remember that this is what the Central Limit Theorem tells us--, we know such facts as that 95% of the samples will have means that fall within a distance of 1.96 standard deviations of the set of sample means on either side of the population mean. When you have drawn a sample of size  $n$ , you have drawn one sample out of the set of all possible samples of that size; and the CLT tells you that your sample has a 95% chance of being among the 95% of all possible samples that lie within a distance of 1.96 standard deviations of the set of sample means.

Putting this into symbols, we can say that your sample has a 95% chance of being within a distance equal to  $1.96\sigma_{\bar{x}(n)}$  (read this as "1.96 standard errors of the mean") from the population mean,  $\mu$ . If you happen to be working with the kind of population for which you know the population standard deviation,  $\sigma$ , then you calculate this distance numerically as

$$1.96\sigma_{\bar{x}(n)} = 1.96\frac{\sigma}{\sqrt{n}}$$

If, on the other hand, you are in the much more common circumstance of not knowing the population standard deviation, then you use the sample standard deviation  $s$  to calculate the distance numerically as

$$1.96s_{\bar{x}(n)} = 1.96\frac{s}{\sqrt{n}}$$

Since the latter situation is the usual one, we shall use the latter formula in what follows.

We have just reviewed why you can trust your observed sample mean,  $\bar{x}$ , to have a 95% chance of being no more than  $1.96s_{\bar{x}(n)}$  ("1.96 standard errors of the mean") away from the population mean,  $\mu$ . Assume for the moment that your sample is indeed one of these "well behaved" samples, the ones that fall within the central 95% interval around  $\mu$ . Then if you go out a distance of  $1.96s_{\bar{x}(n)}$  on either side of your sample mean,  $\bar{x}$ , you will have defined an interval that will contain the population mean. Putting this into the notation we are using, you can make the statement:

$$\text{I am 95\% confident that } \mu \text{ is in the interval } \bar{x} \pm 1.96s_{\bar{x}(n)} = (\bar{x} - 1.96s_{\bar{x}(n)}, \bar{x} + 1.96s_{\bar{x}(n)})$$

By applying the foregoing kind of reasoning to the 99% central interval around the population mean, you can make the following statement:

$$\text{I am 99\% confident that } \mu \text{ is in the interval } \bar{x} \pm 2.58s_{\bar{x}(n)} = (\bar{x} - 2.58s_{\bar{x}(n)}, \bar{x} + 2.58s_{\bar{x}(n)})$$

The above statement uses the value 2.58, which applies to the 99% central intervals of Gaussianly distributed phenomena, in place of the 95% value, 1.96.

## EXAMPLES OF CALCULATING A CONFIDENCE INTERVAL

To illustrate the preceding discussion, we can use the data that Stephens<sup>1</sup> uses in his Example 8.2 (pp. 167-168). He says that a sample of 75 policyholders of an insurance yielded an average age of 30.5 years and that the known standard deviation of the set of ages of policyholders is 5.5 years. From these data he concludes that one can be 95% confident that the average age in years of all policyholders is in the interval

$$\bar{x} \pm 1.96\sigma_{\bar{x}(n)} = 30.5 \pm 1.96\frac{5.5}{\sqrt{75}} = 30.5 \pm 1.96(.635) = 30.5 \pm 1.245 = (29.255, 31.745)$$

As another illustration, we use Stephens's Example 8.3 (p. 168). In this example, Stephens tells us that the response times in a sample of 35 terrorist-bomb threats averaged 8.5 minutes with a sample standard deviation of 4.5 minutes. From these data he concludes that one can be 99% confident that the average response time in minutes to all such threats is in the interval

$$\bar{x} \pm 2.58s_{\bar{x}(n)} = 8.5 \pm 2.58\frac{4.5}{\sqrt{35}} = 8.5 \pm 2.58(.761) = 8.5 \pm 1.961 = (6.539, 10.461)$$

## SMALL VS. LARGE SAMPLES: USING THE STUDENTS' $t$ DISTRIBUTION

When one is dealing with small samples, slight modifications to the foregoing procedures are necessary. What do we mean by "small"? For the purposes of constructing confidence intervals, a widely used rule-of-thumb is that samples of size 31 or more can be considered "large" samples; thus, "small" samples are those of size 30 or below. The distinction arises from the fact that for small samples, the 95% and 99% confidence factors, 1.96 and 2.58, respectively, which come from the Gaussian distribution, need to be replaced by slightly larger confidence factors that come from the Student's  $t$  distribution.

For the construction of confidence intervals, you need to select the  $t$ -distribution value that corresponds to (1) the size of the central interval (95%, 99%, etc.) that you want to use and (2) the appropriate value for the number of degrees of freedom, which will be  $df = n - 1$ .

The Student's  $t$  distribution is actually a family of distributions. They vary according to the size of a parameter known as "degrees of freedom" (usually abbreviated to "df"), but all of them are closely related to the Gaussian distribution. In fact, if you look at a typical table of the Student's  $t$  distribution (e.g., the table on p. 307 of the Hinton text<sup>2</sup>), you will see that the bottom row, which corresponds to an infinite number of degrees of freedom (i.e., for which  $df = \infty$ ), contains Gaussian values including 1.960 and 2.576. You can also see that as you go down a column in the table, the values of  $t$  steadily decrease toward the Gaussian value in the bottom row. In other words, the Students'  $t$  distribution for an infinite number of degrees of freedom is the same as the Gaussian distribution.

Actually, in the construction of confidence intervals, values from the  $t$  distribution will always be slightly more accurate than the Gaussian values, no matter how large the sample is. The rule-of-thumb can be better stated as: "In constructing confidence intervals, it is always preferable to use confidence factors taken from the Student's  $t$  distribution. Hence, for confidence factors for samples of size 31 or more, you are **encouraged** to use  $t$ -distribution values, but for samples of size 30 or smaller, you are **required** to use  $t$ -distribution values."

## EXAMPLES OF CALCULATING A CONFIDENCE INTERVAL USING $t$ -DISTRIBUTION VALUES

In his Example 8.7 (pp. 171-172) Stephens works out the 90% confidence interval for the population mean, based on a sample of 20 observations of vacation-travel distances in hundreds of miles. The sample mean was 12.375 (i.e., 1,237.5 miles), and the sample standard deviation was 3.741. According to our rule-of-thumb, we are required to use the Student's  $t$  distribution for the value of the confidence factor, and Stephens explains why the pertinent value is  $t = 1.729$ . We can work out the 90% confidence interval as follows<sup>3</sup>:

$$\bar{x} \pm 1.729 s_{\bar{x}(n)} = 12.375 \pm 1.729 \frac{3.741}{\sqrt{20}} = 12.375 \pm 1.729(.837) = 12.375 \pm 1.446 = (10.929, 13.821)$$

Now let us work this same problem using *Microsoft Excel*. As is explained in the course Webnote entitled "Notes on Using *Microsoft Excel* for Statistical Analysis," you should enter the data into a column (or row, if you prefer) and then apply the Tools→Data Analysis→Descriptive Statistics procedure. In the Descriptive Statistics procedure, you should enter the spreadsheet range that contains the data, check the Summary Statistics option, and check the Confidence Level for the Mean option, specifying 90% as the level. Here is what the Descriptive Statistics procedure yields for the data in Stephens's Table 8.4:

Mean	12.375
Standard Error	0.8365617
Median	12.25
Mode	12
Standard Deviation	3.7412178
Sample Variance	13.996711
Kurtosis	-0.2965667
Skewness	0.0445488
Range	15
Minimum	5
Maximum	20
Sum	247.5
Count	20
Confidence Level(90.0%)	1.4465251

A good feature of *Excel's* Descriptive Statistics procedure is that it uses  $t$ -distribution values no matter what the size of the sample is; i.e., even for large samples, *Excel* uses the slightly more accurate  $t$  values.

### **Endnotes**

<sup>1</sup> Stephens, L. J. *Schaum's Outline of Theory and Problems of Beginning Statistics*. New York, NY: McGraw-Hill; 1998. ISBN:0-07-061259-5.

<sup>2</sup> Hinton, P. R. *Statistics Explained: A Guide for Social Science Students*. New York, NY: Routledge; 1995. ISBN:0-415-10286-3.

<sup>3</sup> Our values for the 90% confidence interval differ from those of Stephens in the third decimal place; this is undoubtedly due to differences in rounding off.