

Estimation

Xiuxia Du

Department of Bioinformatics & Genomics
University of North Carolina at Charlotte

`Xiuxia.Du@uncc.edu`

Fall 2009

Contents

1	Statistical inference	3
2	Sampling	4
2.1	Random sample	5
2.2	Sample statistic	5
2.3	Location statistics	7
2.4	Measures of variability	8
2.5	Boxplots	9
3	Estimation	14
3.1	Estimator	14
3.2	Sampling distribution	16
3.2.1	Construction of sampling distributions	16
3.2.2	Important characteristics of sampling distributions	16
3.3	Estimation of the mean of a distribution	17
3.3.1	Sampling distribution of the mean	17
3.3.2	Confidence interval	23
3.3.3	t distribution	28
3.4	Estimation of the variance of a distribution	33
3.4.1	Point estimation	33
3.4.2	Interval estimation	35
3.5	Estimation of the difference between two sample means	38
3.5.1	Construction of the sampling distribution of $\bar{x}_1 - \bar{x}_2$	39
3.5.2	Characteristics of the sampling distribution of $\bar{x}_1 - \bar{x}_2$	39
3.5.3	Confidence interval	41

1 Statistical inference

Thus far, we have considered random variables from a theoretical point of view. We have studied two functions, the density and the cumulative distribution function, that enable us to predict the behavior of the variable in a probabilistic sense. We have also considered three parameters that characterize or describe a random variable, namely, μ, σ^2, σ . In practice, the exact distribution of a random variable is seldom known. Rather, we must determine a reasonable form for the density and appropriate values for the density and appropriate values for the distribution parameters from a dataset.

Example 1 (A clinical trial). It is very common for patients with episodes of depression to have a recurrence within two to three years. Prien et al. (1984) studied three treatments for depression: imipramine, lithium carbonate, and a combination. As is traditional in such studies (called *clinical trials*), there was also a group of patients who received a placebo.

In this example, we shall consider 150 patients who entered the study after an episode of depression that was classified as “unipolar” (meaning that there was no manic disorder). They were divided into four groups (three treatments plus placebo) and followed to see how many had recurrences of depression. The following table summarizes the results.

Table 1: Results of clinical depression study

Response	Imipramine	Lithium	Combination	Placebo
Relapse	18	13	22	24
No relapse	22	25	16	10

Question: What can we say about the probability that a patient will respond successfully to treatment after we observe the results from a collection of other patients?

This is the kind of question that statistical inference is designed to address. In general, **statistical inference consists of making probabilistic statements about unknown quantities**. Statistical inference can be further subdivided into the two main areas: estimation and hypothesis testing. **Estimation** is concerned with estimating the values of specific population parameters; **hypothesis testing** is concerned with testing whether the value of a population parameter is equal to some specific value.

Definition 1 (Statistical inference). Statistical inference is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.

In the clinical trial example, we can think of the probability P that a patient in the imipramine group will avoid relapse as an unknown quantity with a probability distribution function describing our uncertainty about P . The patients in the imipramine column should provide us with some information that reduces the uncertainty about P .

Example 2. Suppose we measure the systolic blood pressures of a group of Samoan villagers and we believe the underlying distribution is normal. How can the parameters of this distribution μ, σ^2 be estimated? How precise are our estimates?

Example 3. Suppose we look at people living within a low-income census tract in an urban area and we wish to estimate the prevalence of HIV-positive people in the community. We assume that the number of cases among n people sampled is binomially distributed, with some parameter p . How is the parameter p estimated? How precise is this estimate?

In Examples 2 and 3, we are interested in obtaining specific values as estimates of our parameters. These values are often referred to as **point estimates**. Sometimes we want to specify a range within which the parameter values are likely to fall. If this range is narrow, then we may feel our point estimate is good. This type of problem involves **interval estimation**.

2 Sampling

Often in practice, the group of individuals or objects that we are interested in is quite large. Instead of examining the entire group, called the **population**, which may be difficult or impossible to do, we can examine only a small part of this population, which is called a **sample**. The problem of inferring characteristics of a population from a sample is the central concern of statistical inference. The process of obtaining samples is called **sampling**.

Clearly, the reliability of conclusions drawn concerning a population depends on whether the sample is properly chosen so as to represent the population sufficiently well, and one of the important problems of statistical inference is just how to choose a sample.

One way to do this for finite population is to make sure that each member of the population has the same chance of being in the sample, which is often called a **random sample**. That is, The selection of one object is independent of the selection of any other. Random sampling can be accomplished for relatively small populations by using a table of random numbers specially constructed for such purposes.

After a sample is drawn, we must devise methods for approximating the characteristics of a population based on the sample.

Example 4. We may wish to draw conclusions about the heights of 12,000 students (the population) by examining only 100 students (a sample) selected from the population.

2.1 Random sample

Note that, prior to the actual selection of the students to be studied, $X_i (i = 1, 2, \dots, 100)$, the heights of the i th student selected is a random variable. It has the same distribution as X , the height of student in the population. Furthermore, these random variables are independent in the sense that the values assumed by one has no effect on the value assumed by any of the others. The random variables X_1, X_2, \dots, X_{100} can be thought of as a random sample.

The term **random sample** is often used in three different but closely related ways in applied statistics. It may refer to (a) the **objects** selected for study, (b) to the **random variables** associated with the objects to be selected, (c) or to the **numerical values** assumed by those variables. It is usually clear from the context which is intended.

Definition 2 (Random sample). A random sample of size n from the distribution of X is a collection of n independent random variables, each with the same distribution as X .

2.2 Sample statistic

When objects are selected from a finite population, this type of sample results only when sampling is done with replacement. This ensures that X_1, X_2, \dots, X_n are indeed i.i.d. Usually, sampling from a finite population is done **without** replacement. This means that the random variables X_1, X_2, \dots, X_n are NOT independent. However, if the sample is small relative to the population itself, then removal of a few items does not drastically alter the composition of the population. A generally accepted guideline is that for all practical purposes we may assume independence whenever the sample constitutes at most 5% of the population. If this is not true, then the techniques used to estimate parameters must be altered to take this into account.

Once a sample has been drawn, we commonly use the data gathered to evaluate pertinent **statistics**. What is a statistic? Roughly speaking, a statistic is a random variable whose numerical value can be determined from a random sample. That is, a statistic is a random variable that is a function of the elements of a random sample X_1, X_2, \dots, X_n . Typical statistics of interest to statisticians are

$$\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i/n, \max_i(X_i), \min_i(X_i)$$

Example 5. Consider the random variable X , the number of times per hour that a television signal is interrupted by random interference. Assume that this random variable has a Poisson distribution with unknown mean μ and unknown variance σ^2 . To approximate the value of each of these parameters, we intend to observe the signal for ten randomly selected nonoverlapping one-hour periods over a week's time.

Let $X_i, i = 1, 2, \dots, 10$ denote the number of interruptions that occur during the i th observation period. The random variables X_1, X_2, \dots, X_n constitute a random sample of size 10 from a Poisson distribution. When the experiment is conducted, these data result:

$$\begin{aligned} x_1 = 1 \quad x_2 = 0 \quad x_3 = 1 \quad x_4 = 0 \quad x_5 = 3 \\ x_6 = 0 \quad x_7 = 2 \quad x_8 = 1 \quad x_9 = 0 \quad x_{10} = 0 \end{aligned}$$

The observed values of the statistics $\sum_{i=1}^n X_i$, $\sum_{i=1}^n X_i^2$, $\sum_{i=1}^n X_i/n$, $\max_i(X_i)$, $\min_i(X_i)$ based on this sample are 8, 16, .8, 3, and 0, respectively. Note that the random variable $X_1 - \mu$ is **not** a statistic. Since μ is unknown, we cannot determine its numerical value from a random sample.

Generally, any quantity obtained from a sample for the purpose of estimating a population parameter is called a **sample statistic**, or briefly **statistic**.

Mathematically, a sample statistic for a sample of size n can be defined as a function of the random variables X_1, X_2, \dots, X_n

$$g(X_1, X_2, \dots, X_n)$$

This function is another random variable, whose values can be represented by

$$g(x_1, x_2, \dots, x_n)$$

The word **statistic** is often used for the random variable or for its values, the particular sense being clear from the context.

Next, we consider some statistics that allow us to summarize a dataset analytically. Since it is hoped that the dataset reflects the population as a whole, these statistics also give us some idea of the values of the parameters that characterize X over the population under study. In particular, we consider measures of location or central tendency in a dataset, the **sample mean**, and the **sample median**. We also consider measures of variability within the data set, the **sample variance**, **sample standard deviation**, and the **sample range**. The word **sample** is used to emphasize the fact that they represent a random sample from the distribution of X .

The probability distribution of a sample statistic is called the **sampling distribution** of the statistic. For a sampling distribution, we can compute a mean, variance, standard deviation, moments, etc. The standard deviation is sometimes also called the **standard error**.

2.3 Location statistics

Definition 3 (Sample mean). Let X_1, X_2, \dots, X_n be a random sample from the distribution of X . Then the **mean of the sample** or **sample mean** is a random variable defined by

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

This definition of sample mean is also called the arithmetic mean in statistics.

If x_1, x_2, \dots, x_n denote values obtained in a particular sample of size n , then the mean for that sample is denoted by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Note that μ_X and \bar{X} are **not** the same. The parameter μ_X is the theoretical average value for X over the entire population. \bar{X} is a statistic which, when evaluated over a particular random sample, gives the average value of X **for that sample**. It is hoped, of course, that the observed value of \bar{X} is close to μ_X .

Definition 4 (Sample median). Let x_1, x_2, \dots, x_n be a sample of observations arranged in order from the smallest to the largest. The sample median is the middle observation if n is odd. It is the average of the two middle observations if n is even.

$$\text{Median location} = \frac{n+1}{2}$$

There are an equal number of sample points on both sides of the sample median.

Definition 5 (Mode). The mode of a probability distribution for X is the value x at which its probability mass function (for discrete random variable) or probability density function (for continuous random variable) attains its maximum value.

Clearly, in statistics, the mode is the value that occurs the most frequently in a data set.

Definition 6 (Geometric mean). The geometric mean of a sequence of data points $x_i, i = 1, 2, \dots, n$ is defined as

$$\sqrt[n]{x_1 x_2 \dots x_n}$$

Many types of laboratory data, specifically data in the form of concentrations of one substance in another, as assessed by serial dilution techniques, can be expressed either as multiples of 2 or as a constant multiplied by a power 2. The arithmetic mean is not appropriate as a measure of location in this situation, because the distribution is skewed. One way to work with this type of data is to take the logarithmic transformation of the

data, study the data in the logarithmic space and finally transform the data back to the normal space. Average can be taken in the logarithmic space:

$$\frac{1}{n} (\log_2 x_1 + \log_2 x_2 + \cdots + \log_2 x_n)$$

The anti-logarithm of the above average is then the geometric mean.

2.4 Measures of variability

Definition 7 (Sample Variance and sample standard deviation). Let X_1, X_2, \dots, X_n be a random sample of size n from the distribution of X . Then the statistic

$$S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1} \quad (1)$$

is called the **sample variance**. Furthermore, the statistic $S = \sqrt{S^2}$ is called the **sample standard deviation**.

Theorem 1 (A computational formula for S^2). Let X_1, X_2, \dots, X_n be a random sample of size n from the distribution of X . The sample variance is given by

$$S^2 = \frac{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2}{n(n-1)}$$

Word of caution: We have assumed that the data set presented so far represents a random sample drawn from a larger population because this is the situation most often encountered in practice. Occasionally, you will encounter a data set that is **not** a sample. Rather, it represents an observation on X for **every** member of the population. If this is the case, then the population mean is just the arithmetic average of these observations; that is, $\mu = \bar{x}$. Furthermore, the population variance is given by

$$\sigma^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n}$$

Be careful! Be sure that you understand the nature of your data before you begin to summarize its properties.

Definition 8 (Range). The range is the difference between the largest and smallest observations in a sample.

Definition 9 (Quantile). For a population of discrete values or for a continuous population density, the k th q -quantile is the data value where the cumulative distribution function crosses k/q . That is x is a k th q -quantile for a random variable X if

$$P[X \leq x] \leq \frac{k}{q}$$

For a finite population of N values indexed from 1 to N from lowest to highest, the k th q -quantile of this population can be computed via the value $I = N \frac{k}{q}$. If I is not an integer, then round up to the next integer to get the appropriate index, the corresponding data value is the k th q -quantile. On the other hand, if I is an integer, then any number from the data value at that index to the data value of the next can be taken as the quantile, and it is conventional to take the average of those two values.

The 100-quantile is called **percentile**. The median, being the 50th percentile, is a special case of a quantile. The 4-quantile is called a **quartile**. That is, a quartile is any of the three values which divide the sorted data set into four equal parts, so that each part represents one fourth of the sample data.

It is useful to relate the arithmetic mean and the standard deviation to each other, because, for example, a standard deviation of 10 means something different conceptually if the arithmetic mean is 10 than if it is 1000. A special measure, the coefficient of variation, is often used for this purpose.

Definition 10 (Coefficient of variation (CV)). The coefficient of variation is defined as the ratio of the standard deviation σ to the mean μ :

$$c = \frac{\sigma}{\mu}$$

This measure remains the same regardless of what units are used, because if the units change by a factor, both the mean and standard deviation change by the same factor. CV, which is the ratio between them, remains unchanged [2].

2.5 Boxplots

In summarizing data, it is useful to report all the statistics. This is especially true if the dataset contains a value that is unusually large or unusually small. A value that appears to be atypical in that it seems to be far removed from the bulk of the data is called an **outlier**. It is important to be able to detect such numbers and to understand the effect that they have on the usual sample statistics.

Outliers arise for two reasons: (1) They are legitimate observations whose values are simple unusually large or unusually small, or (2) they are the result of an error in measurement, poor experimental technique, or a mistake in recording or entering the

data. In the first case, it is suggested that the presence of the outlier be reported and that sample statistics be reported both with and without the outlier. In the second case, the data point can be corrected if possible or else dropped from the data set.

Of the statistics presented thus far, the sample mean, the variance, the standard deviation, and the range are adversely affected by the presence of an outlier; however, the sample median is not so affected. Thus in the presence of an outlier the sample median may be preferable to the sample mean measure of location. We say that the median is resistant to outliers.

Sometimes, outliers are so obvious that their presence can be detected by inspection. However, it is useful to have an analytical and graphical technique for identifying values that are truly unusual. One such technique is the **boxplot**. Its construction is based on the interquartile range, a measure of variability that is resistant to outliers. The sample interquartile range, *iqr*, represents the length of the interval that contains roughly the middle 50% of the data. If the *iqr* is small, then much of the data lies close to the center of the distribution. If it is large, the data tend to be widely dispersed.

Finding the Sample Interquartile range

1. Find the median location $(n + 1)/2$, where n is the sample size.
2. Truncate the median location by rounding it **down** to the nearest whole number
3. Find the quartile location q by

$$q = \frac{\text{truncated median location} + 1}{2}$$

4. Find q_1 by counting up from the smallest data point to location q . If q is an integer, then q_1 is the data point in position q . If q is not an integer, then q_1 is the average of the data points in positions $q - 0.5$ and $q + 0.5$. Approximately 25% of the data will fall on or below q_1 .
5. Find q_3 by counting down from the largest data point to position q as in part 4. Approximately 75% of the data will fall on or below q_3 .
6. Define *iqr* by $\text{iqr} = q_3 - q_1$.

Once the interquartile range has been found, it can be used to construct a boxplot. The boxplot is a graphical representation of a data set that gives a visual impression of location, spread, and the degree and direction of skewness. For approximately bell-shaped distributions, the boxplot also allows us to identify outliers. It is especially useful when we want to compare two or more data sets.

Constructing a boxplot

1. A horizontal or vertical reference is constructed.
2. Find the sample median, q_1 , q_3 , and iqr.
3. Find two points f_1 and f_3 , called **inner fences**, by

$$f_1 = q_1 - 1.5(\text{iqr})$$

$$f_2 = q_3 + 1.5(\text{iqr})$$

These points will be used to identify outliers.

4. Find two points a_1 and a_3 , called **adjacent values**. The point a_1 is the data point that is closest to f_1 without lying below f_1 in value. The point a_3 is the data point that is closest to f_3 without lying above f_3 in value.
5. Find two points F_1 and F_3 , called **outer fences**, by

$$F_1 = q_1 - 2(1.5)(\text{iqr})$$

$$F_2 = q_3 + 2(1.5)(\text{iqr})$$

6. Locate the points found thus far on the horizontal or vertical scale. Their relative positions are shown in Fig.
7. Construct a box with ends at q_1 and q_3 with an interior line drawn at the median as shown in Fig.
8. Indicate adjacent values by x and connect them to the box with dashed lines. Locate any points falling between the inner and outer fences and denote these by open circles. These points are considered to be mild outliers. Indicate data points that fall beyond the outer fences with asterisks. These points are considered to be extreme outliers.

The location of the midline of the box is an indication of the shape of the distribution. If the line is badly off center, then we know that the distribution is skewed in the direction of the longer end of the box.

Before we illustrate this technique, the notion of fences needs to be clarified. It can be shown that when sampling from a normal distribution, only about 7 values in every 1000 fall beyond the inner fences. Since these values are very unusual, they are deemed to be outliers. Outliers must be treated with care since their presence can have a dramatic impact on \bar{x} , s^2 , s , the usual measures of location and variation. When an outlier is found, we should consider its source. Is it a legitimate data point whose value is simply unusually large or small? Is it a misrecorded value? Is it the result of some

error or accident in experimentation? In the last two instances, the point can be deleted from the data set and the analysis completed on the remaining data. In the first case it is suggested that the presence of the outlier be made known and that statistics be reported both with and without the outlier [1].

Example 6. A sample consists of the following data points:

10, 21, 12, 12, 20, 13, 24, 36, 31, 18, 17, 16, 37, 16, 32, 13, 14, 49, 25, 19, 13, 32, 27

Construct the boxplot for this sample.

Solution:

Sort the data points, we get

10, 12, 12, 13, 13, 13, 14, 16, 16, 17, 18, 19, 20, 21, 24, 25, 27, 31, 32, 32, 36, 37, 49

The total number of data points is $n = 23$.

1. The median location is $= \frac{n+1}{2} = \frac{23+1}{2} = 12$

2. The location is already a whole number, no need to truncate.

3. Find quartile q :

$$q = \frac{12+1}{2} = 6.5$$

4. Find q_1 . q is not an integer, then q_1 is the average of the data points in positions $q - 0.5 = 6$ and $q + 0.5 = 7$, that is

$$q_1 = \frac{13+14}{2} = 13.5$$

5. Find q_3 . q_3 is the average of the data points in positions $n - 6 + 1 = 23 - 6 + 1 = 18$ and $n - 7 + 1 = 23 - 7 + 1 = 17$, that is,

$$q_3 = \frac{27+31}{2} = 29$$

6. $\text{iqr} = q_3 - q_1 = 29 - 13.5 = 15.5$

7. Find inner fences as follows:

$$f_1 = q_1 - 1.5(\text{iqr}) = 13.5 - 1.5(15.5) = -9.75$$

$$f_3 = q_3 + 1.5(\text{iqr}) = 29 + 1.5(15.5) = 52.25$$

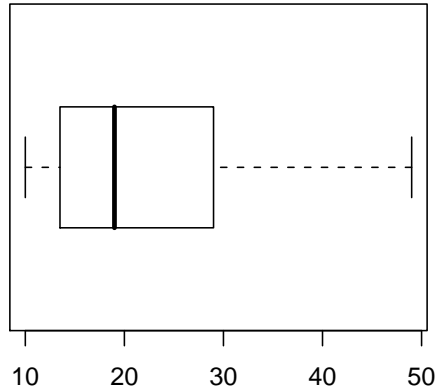


Figure 1: Boxplot for example 6

8. Find a_1 and a_3 :

$$a_1 = 10$$

$$a_3 = 49$$

Fig. 1 shows the boxplot.

Example 7. A study of posttraumatic amnesia after a closed head injury is conducted. One variable studied is the length of hospitalization in days. The collected data is

8, 12, 20, 27, 30, 32, 35, 36, 40, 40, 40, 40, 41, 42, 45, 47, 50, 52, 61, 89, 108

Construct the boxplot for this sample.

Solution: The sample size is $n = 21$. Sort the data:

8, 12, 20, 27, 30, 32, 35, 36, 40, 40, 40, 40, 41, 42, 45, 47, 50, 52, 61, 89, 108

1. The median location is $= \frac{n+1}{2} = \frac{21+1}{2} = 11$ and the median is 40.
2. The location is already a whole number, so no need to truncate.

3. Find quartile q :

$$q = \frac{11 + 1}{2} = 6$$

4. Find q_1 which is the data point in position q . Thus $q_1 = 32$.

5. Find q_3 which is the data point in position $n - q + 1 = 21 - 6 + 1 = 16$. Thus $q_3 = 47$.

6. $iqr = q_3 - q_1 = 47 - 32 = 15$.

7. Find inner fences as follows:

$$f_1 = q_1 - 1.5(iqr) = 32 - 1.5(15) = 9.5$$

$$f_3 = q_3 + 1.5(iqr) = 47 + 1.5(15) = 69.5$$

8. The adjacent values are $a_1 = 12$ and $a_3 = 61$.

9. The outer fences are:

$$F_1 = q_1 - 2(1.5)(iqr) = 32 - 2(1.5)(15) = -13$$

$$F_3 = q_3 + 2(1.5)(iqr) = 47 + 2(1.5)(15) = 92$$

Fig. 2 shows the boxplot.

3 Estimation

We have learned how to define sample statistics that allow us to estimate the mean, variance, and standard deviation of a random sample in a logical manner. However, we are unable to assess their effectiveness. Next, we consider the mathematical properties of these and other statistics.

3.1 Estimator

In an estimation problem, there is at least one parameter θ whose value is to be approximated on the basis of a sample. The approximation is done by using an appropriate statistic. A statistic used to approximate or estimate a population parameter θ is called a **point estimator** for θ and is denoted by $\hat{\theta}$; the numerical value assumed by this statistic when evaluated for a given sample is called a **point estimate** for θ . Note the difference between the terms “estimator” and “estimate.” The estimator is the statistic used to generate the estimate and it is a random variable. An estimate is a number.

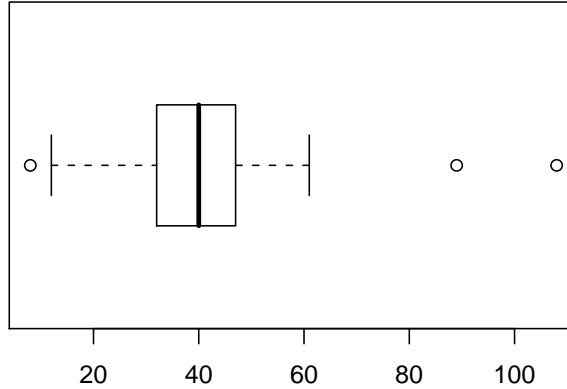


Figure 2: Boxplot for example 7

Once a logical point estimator for a parameter θ has been developed, the natural question to ask is: “How good is the estimator?” Obviously, we want the estimator to generate estimates that can be expected to be close in value to θ . This can be expected to occur if the estimator $\hat{\theta}$ possesses two properties. In particular, we would like

1. $\hat{\theta}$ to be unbiased for θ
2. $\hat{\theta}$ to have a small variance for large sample sizes.

The word “unbiased” means “centered at the right spot” where the right spot is the parameter being estimated.

Definition 11 (Unbiased). An estimator $\hat{\theta}$ is an unbiased estimator for a parameter θ if and only if $E[\hat{\theta}] = \theta$.

Recall that $\hat{\theta}$ is a statistic; therefore it is also a random variable and, as such, has a mean. To say that $\hat{\theta}$ is unbiased for θ implies that the mean of the estimator $\hat{\theta}$ is equal to the parameter θ that is it estimating. Thus an estimator $\hat{\mu}$ is an unbiased estimator for μ if and only if $E[\hat{\mu}] = \mu$; an estimator $\hat{\sigma}^2$ is unbiased for σ^2 if and only if $E[\hat{\sigma}^2] = \sigma^2$; an estimator $\hat{\sigma}$ is unbiased for σ if and only if $E[\hat{\sigma}] = \sigma$. Let us re-examine the estimators \bar{X} , S^2 and S developed in section 2 in light of this new definition.

3.2 Sampling distribution

The importance of a clear understanding of sampling distribution cannot be overemphasized, as this concept is the very key to the understanding of statistical inference. Sampling distributions serve two purposes: (1) they allow us to answer probability questions about sample statistics, and (2) they provide the necessary theory for making statistical inference procedures valid.

Definition 12. The distribution of all possible values that can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population, is called the **sampling distribution** of that statistic.

3.2.1 Construction of sampling distributions

To construct a sampling distribution, we proceed as follows:

1. From a finite population of size N , randomly draw all possible samples of size n .
2. Compute the statistic of interest for each sample.
3. Plot the histogram of the statistic obtained above.

The actual construction of a sampling distribution is a formidable undertaking if the population is of any appreciable size and is an impossible task if the population is infinite. In such cases, sampling distributions may be approximated by taking a large number of samples of a given size.

3.2.2 Important characteristics of sampling distributions

We usually are interested in knowing three things about a given sampling distribution: its **mean**, its **variance**, and its **functional form** (how it looks when graphed).

We can recognize the difficulty of constructing a sampling distribution according to the steps given above when the population is large. We also run into a problem when considering the construction of a sampling distribution when the population is infinite. The best we can do experimentally in this case is to approximate the sampling distribution of a statistic.

Both of these problems may be obviated by means of mathematics. In the sections that follow, some of the more frequently encountered sampling distributions are discussed.

3.3 Estimation of the mean of a distribution

3.3.1 Sampling distribution of the mean

Definition 13 (Sampling distribution of mean). The sampling distribution of \bar{X} is the distribution of values of \bar{x} over all possible samples of size n that could have been selected from the reference population.

Example 8. Suppose we have a population of size $N = 5$, consisting of the ages of five children who are outpatients in a community mental health center. The ages are as follows:

$$x_1 = 6, x_2 = 8, x_3 = 10, x_4 = 12, x_5 = 14$$

The mean, μ , of this population is equal to $(\sum x_i)/N = 10$ and the variance is

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{40}{5} = 8$$

We wish to construct the sampling distribution of the sample mean, \bar{x} , based on samples of size $n = 2$ drawn from this population.

Solution: Let us draw all possible samples of size $n = 2$ from this population. These samples, along with their means, are in Figure (3.3.1). The histograms are shown in Figure (4). We see in this example that, when sampling is with replacement, there are

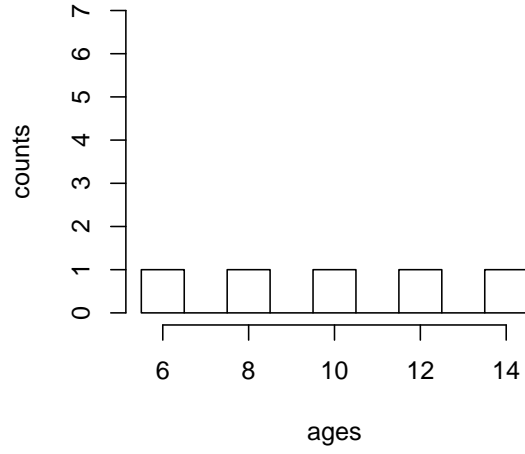
		Second Draw				
		6	8	10	12	14
First Draw	6	(6,6)	(6,8)	(6,10)	(6,12)	(6,14)
	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	
	8	(8,6)	(8,8)	(8,10)	(8,12)	(8,14)
	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	
	10	(10,6)	(10,8)	(10,10)	(10,12)	(10,14)
<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>		
12	(12,6)	(12,8)	(12,10)	(12,12)	(12,14)	
<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>		
14	(14,6)	(14,8)	(14,10)	(14,12)	(14,14)	
<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>	<u>14</u>		

Figure 3: All possible samples of size $n = 2$ from a population of size $N = 5$. Samples above or below the principal diagonal result when sampling is without replacement. Sample means are in red.

25 possible samples. In general, when sampling is with replacement, the number of possible samples is equal to N^2 .

It was stated earlier that we are usually interested in the functional form of a sampling distribution, its mean, and its variance. We now consider these characteristics for the sampling distribution of the sample mean, \bar{x} .

(A)



(B)

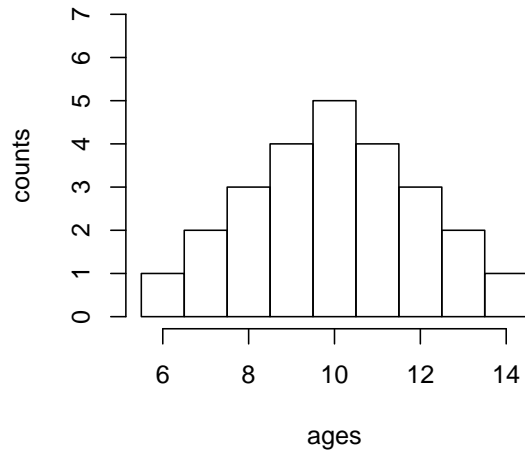


Figure 4: Distribution of population (A) and sampling distribution of \bar{X} (B).

Example 9. Sampling distribution of mean for the birth weight population.

Theorem 2. Let X_1, X_2, \dots, X_n be a random sample of size n from a distribution with mean μ . The sample mean, \bar{X} , is an unbiased estimator for μ .

Proof.

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] \\ &= \frac{1}{n}(E[X_1] + E[X_2] + \dots + E[X_n]) \end{aligned}$$

Since X_1, X_2, \dots, X_n constitutes a random sample from a distribution with mean μ , each of these random variables has mean μ . Therefore,

$$E[\bar{X}] = \frac{1}{n}(n\mu) = \mu$$

□

It is important to realize that since $\hat{\theta}$ is a statistic, in repeated sampling the estimates generated will vary from sample to sample. To say that $\hat{\theta}$ is unbiased for θ implies that these estimates vary about θ ; it also implies that the **average** value of these estimates can be expected to lie reasonably close to θ . For example, since \bar{X} is unbiased for μ , for k repetitions of an experiment the observed sample means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ will vary about μ and the **average** value of these k estimates should lie reasonably close to μ .

Example 10. The birth weights from 1000 consecutive deliveries at Boston City Hospital are enumerated. For this example, consider this population as effectively infinite. Suppose we wish to draw a sample of size 10 from this population and get the following data points (in oz):

97, 117, 140, 78, 99, 148, 108, 135, 126, 121

What is the average birth weight for this sample?

Solution:

$$\bar{x} = \frac{97 + 117 + 140 + 78 + 99 + 148 + 108 + 135 + 126 + 121}{10} = 116.9$$

It is equally important to understand what the term “unbiased” does **not** imply. It does not imply that any **one** estimate will be close in value to the parameter being estimated. In Example 10, the estimated mean birth weight is $\hat{\mu} = 116.9$. This estimate

is unbiased in the sense that it was generated by means of the unbiased estimator \bar{X} . This **alone** does not guarantee that the actual mean birth weight of **all the 1000 babies** is anywhere close to 116.9 oz. This is unfortunate. Usually, statistical studies are not repeated over and over so that the estimates obtained can be averaged. In general, only one sample is drawn, one estimate is obtained. To have some assurance that this estimate is close in value to θ , the parameter being estimated, ideally the estimator used not only should be unbiased, but also it should have a small variance for large sample sizes. In this way, even though the estimated values fluctuate about θ , the variability is small. Each estimate produced can be expected to be fairly close in value to θ . The following theorem shows that \bar{X} has this property.

Theorem 3. Let \bar{X} be the sample mean based on a random sample of size n from a distribution with mean μ and variance σ^2 . Then,

$$Var[\bar{X}] = \frac{\sigma^2}{n}$$

Proof.

$$\begin{aligned} Var[\bar{X}] &= Var\left[\frac{1}{n}\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n^2}Var\left[\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n Var[X_i] \\ &= \frac{1}{n^2}\sum_{i=1}^n \sigma^2 \\ &= \frac{1}{n^2}(n\sigma^2) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

□

Note that since σ^2 is constant, as the sample size n increases, the variance of \bar{X} , σ^2/n , decreases and can be made as small as we wish by choosing n sufficiently large. This implies that a sample mean based on a large sample can be expected to lie reasonably close to μ ; one based on a small sample may vary widely from the actual population mean. This points out the advantages of working with a large sample and the danger of placing too much emphasis on conclusions drawn from small samples.

Example 11 (Illustration of $Var(\bar{X})$ vs. sample size n). Birth weight example.

The unbiasedness of \bar{X} is not sufficient reason to use it as an estimator of μ . Many unbiased estimators of μ exist, including the sample median and the average value of the largest and smallest data points in a sample. Why is \bar{X} chosen rather than any of the other unbiased estimators? The reason is that if the underlying distribution of the population is normal, then it can be shown that the unbiased estimator with the smallest variance is given by \bar{X} . Thus, \bar{X} is called the **minimum variance unbiased estimator** of μ .

Example 12 (Illustration of the minimum variance unbiased estimator). Birth weight example. Sampling distribution of the sample mean, sample median, and the average of the smallest and largest observations.

Since the standard deviation of any random variable is the square root of its variance, the standard deviation of the sample mean is the square root of the variance of \bar{X} . Thus, the standard deviation of \bar{X} is σ/\sqrt{n} and is referred to as the **standard error the mean** or the **standard error**.

Theorem 4 (Distribution of \bar{X} - normal distribution). Let X_1, X_2, \dots, X_n be a random sample of size n from a normal distribution with mean μ and variance σ^2 . Then \bar{X} is normally distributed with mean μ and variance σ^2/n .

Theorem 5 (Central limit theorem for the distribution of \bar{X}). Let X_1, X_2, \dots, X_n be a random sample of size n from a distribution with mean μ and variance σ^2 . Then for large n , \bar{X} is approximately normal with mean μ and variance σ^2/n . Furthermore, for large n , the random variable $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is approximately standard normal.

Example 13 (Illustration of the central limit theorem). Birth weight example. Distribution of \bar{X} vs sample size n .

A mathematical formulation of the central limit theorem is that the distribution of

$$\frac{\bar{x} - \mu}{\sigma\sqrt{n}}$$

approaches a normal distribution with mean 0 and variance 1 as $n \rightarrow \infty$. Note that the central limit theorem allows us to sample from nonnormally distributed populations with a guarantee of approximately the same results as would be obtained if the populations were normally distributed provided that we take a large sample.

The importance of this will become evident later when we learn that a normally distributed sampling distribution is a powerful tool in statistical inference. In the case of the sample mean, we are assured of at least an approximately normally distributed

sampling distribution under three conditions: (1) when sampling is from a normally distributed population; (2) when sampling is from a nonnormally distributed population and our sample is large; and (3) when sampling is from a population whose functional form is unknown to us as long as our sample size is large.

The logical question that arises at this point is, how large does the sample have to be in order for the central limit theorem to apply? There is no one answer, since the size of the sample needed depends on the extent of nonnormality present in the population. One rule of thumb states that, in most practical situations, a sample of size 30 is satisfactory. In general, the approximation to normality of the sampling distribution of \bar{x} becomes better and better as the sample size increases.

Sampling without replacement: The forgoing results have been given on the assumption that sampling is either with replacement or that the samples are drawn from infinite populations. In general, we do not sample with replacement, and in most practical situations it is necessary to sample from a finite population. Hence, we need to become familiar with the behavior of the sampling distribution of the sample mean under these conditions.

Theorem 6. When sampling is without replacement from a finite population, the sampling distribution of \bar{x} will have mean μ and variance

$$\sigma_{\bar{x}^2} = \frac{\sigma^2}{n} \frac{N-n}{N-1}$$

If the sample size is large, the central limit theorem applies and the sampling distribution of \bar{x} will be approximately normally distributed.

The finite population correction The factor $(N-n)/(N-1)$ is called the finite population correction and can be ignored when the sample size is small in comparison with the population size. When the population is much larger than the sample, the difference between σ^2/n and $(\sigma^2/n)[(N-n)/(N-1)]$ will be negligible. Imagine a population of size 10,000 and a sample from this population of size 25; the finite population correction would be equal to $(10000-25)/9999=.9976$. Most practicing statisticians do not use the finite population correction unless the sample is more than 5% of the size of the population. That is, the finite population correction is usually ignored when $n/N \leq .05$.

Summary of the sampling distribution of \bar{x} .

1. Sampling is from a normally distributed population with a known population variance:
 - $\mu_{\bar{x}} = \mu$
 - $\sigma_{\bar{x}} = \sigma/\sqrt{n}$

- The sampling distribution of \bar{x} is normal.
2. Sampling is from a nonnormally distributed population with a known population variance:
- $\mu_{\bar{x}} = \mu$
 - $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, when $n/N \leq .05$
 - $\sigma_{\bar{x}} = (\sigma/\sqrt{n})\sqrt{\frac{N-n}{N-1}}$, otherwise.
 - The sampling distribution of \bar{x} is approximately normal.

Example 14. Compute the probability that the mean birthweight from a sample of 10 infants from the Boston City Hospital population will fall between 98.0 and 126.0 oz if the mean birth weight for the 1000 birth weights from the Boston City Hospital Population is 112.0 oz with a standard deviation of 20.6 oz.

Solution: \bar{X} follows a normal distribution with mean $\mu = 112.0$ and standard deviation $\sigma/\sqrt{n} = 20.7/\sqrt{10} = 6.51$ oz. Let

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 112.0}{6.51}$$

Then, it follows that

$$\begin{aligned} P[98.0 \leq \bar{X} \leq 126.0] &= P\left[\frac{98.0 - 112.0}{6.51} \leq Z \leq \frac{126.0 - 112.0}{6.51}\right] \\ &= P[-2.15 \leq Z \leq 2.15] = .968 \end{aligned}$$

Thus, 96.8% of the samples of size 10 would be expected to have mean birthweights between 98.0 and 126.0 oz.

3.3.2 Confidence interval

Suppose researchers wish to estimate the mean of some normally distributed population. They draw a random sample of size n from the population and compute \bar{x} , which they use as a point estimate of μ . Although this estimator of μ possesses all the qualities of a good estimator, we know that because random sampling inherently involves chance, \bar{x} cannot be expected to be equal to μ .

It would be much more meaningful, therefore, to estimate μ by an interval that somehow communicates information regarding the probable magnitude of μ .

Example 15. Suppose that the first sample of 10 birth weights has been drawn:

97, 117, 140, 78, 99, 148, 108, 135, 126, 121

Our best estimate of the population mean μ would be the sample mean $\bar{x} = 116.9$ oz. Although 116.9 oz is our best estimate of μ , we still are not certain that μ is 116.9 oz. Indeed, if the second sample had been drawn with following data points:

177, 198, 107, 99, 104, 121, 148, 133, 126, 115

a point estimate of 132.8 oz would have been obtained. Our point estimate would certainly have a different meaning if it was highly likely that μ was within 1 oz of 116.9 oz rather than within 1 lb (16 oz).

Point estimation does not give us the ability to report the accuracy of our estimate. To do this, we must turn to the method of interval estimation. The statistics used to extend a point estimate for a parameter θ to an interval of values that should contain the true value of θ vary from parameter to parameter. However, the method for deriving these statistics is basically the same in each case.

Definition 14 (Confidence interval). A $100(1 - \alpha)\%$ confidence interval for a parameter θ is a random interval $[L_1, L_2]$ such that

$$P[L_1 \leq \theta \leq L_2] \doteq 1 - \alpha$$

regardless of the value of θ .

To construct a $100(1 - \alpha)\%$ confidence interval for a parameter θ , we shall find a random variable whose expression involves θ and whose probability distribution is known at least approximately.

Example 16. Acute myeloblastic leukemia is among the most deadly of cancers. Past experience indicates that the time in months that a patient survives after initial diagnosis of the disease is normally distributed with a mean of 13 months and a standard deviation of 3 months. A new treatment is being investigated which should prolong the average survival time without affecting variability. Let X_1, X_2, \dots, X_n denote a random sample from the distribution of X , the survival time under the new treatment. We are assuming that X is normally distributed with $\sigma^2 = 9$ and μ unknown. We want to find statistics L_1 and L_2 so that $P[L_1 \leq \mu \leq L_2] \doteq .95$.

Solution: The random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is standard normal. Thus

$$\begin{aligned} P[-1.96 \leq Z \leq 1.96] &= .95 \\ \Rightarrow P\left[-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right] &= .95 \\ \Rightarrow P\left[-1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}}\right] &= .95 \\ \Rightarrow P\left[-\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right] &= .95 \\ \Rightarrow P\left[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right] &= .95 \end{aligned}$$

From this we see that the lower and upper bounds for a 95% confidence interval are

$$L_1 = \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \quad L_2 = \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

These statistics have the property that in repeated sampling from the population, 95% of the numerical intervals generated are expected to contain μ ; by chance, 5% will not.

Interval estimate components Let us examine the composition of the interval estimate constructed in the example above. It contains in its center the point estimate of μ . The 1.96 we recognized as a value from the standard normal distribution that tells us within how many standard errors lie approximately 95% of the possible values of \bar{x} . This value of z is referred to as the **reliability coefficient**. The last component, $\sigma_{\bar{x}}$, is the standard error, or standard deviation of the sampling distribution of \bar{x} . In general, then, an interval estimate may be expressed as follows:

$$\text{estimator} \pm (\text{reliability coefficient}) \times (\text{standard error})$$

In particular, when sampling is from a normal distribution with known variance, an interval estimate for μ may be expressed as

$$\bar{x} \pm z_{1-\alpha/2} \sigma_{\bar{x}} \tag{2}$$

where $z_{1-\alpha/2}$ is the value of z to the left of which lies $1 - \alpha/2$ and to the right of which lies $\alpha/2$ of the area under its curve.

Interpreting confidence intervals How do we interpret the interval given by Equation (2)? In the present example, where the reliability coefficient is equal to 1.96, we say that in repeated sampling, approximately 95% of the intervals constructed by Equation (2) will include the population mean. This interpretation is based on the probability of occurrence of different values of \bar{x} . We may generalize this interpretation if we designate the total area under the curve of \bar{x} that is outside the interval $\mu \pm 1.96\sigma_{\bar{x}}$ as α and the area within the interval as $1 - \alpha$ and give the following **probabilistic interpretation** of Equation (2).

Probabilistic interpretation

In repeated sampling, from a normally distributed population with a known standard deviation, $100(1 - \alpha)$ percent of all intervals of the form $\bar{x} \pm z_{1-\alpha/2}\sigma_{\bar{x}}$ will in the long run include the population mean μ .

The quantity $1 - \alpha$, in this case, is called the **confident coefficient** (or confidence interval), and the interval $\bar{x} \pm z_{1-\alpha/2}\sigma_{\bar{x}}$ is called a **confidence interval** for μ .

Example 17. Suppose a researcher, interested in obtaining an estimate of the average level of some enzyme in a certain human population, takes a sample of 10 individuals, determines the level of the enzyme in each, and computes a sample mean of $\bar{x} = 22$. Suppose further it is known that the variable of interest is approximately normally distributed with a variance 45. We wish to estimate μ .

Solution: An approximate 95% confidence interval for μ is given by

$$\begin{aligned} & \bar{x} \pm 1.96\sigma_{\bar{x}} \\ & 22 \pm 1.96\sqrt{\frac{45}{10}} \\ & 22 \pm 1.96(2.1213) \\ & [L_1, L_2] = [17.84, 26.15] \end{aligned}$$

In this example we say that we are 95% confident that the population mean is between 17.84 and 26.15. This is called the **practical interpretation** of Equation (2). In general, it may be expressed as follows:

Practical interpretation

When sampling is from a normally distributed population with known standard deviation, we are $100(1 - \alpha)$ percent confident that the single computed interval, $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$, contains the population mean μ .

Researchers may use any confidence coefficient they wish. The most frequently used values are .90, .95, and .99, which have associated reliability factors, respectively, of 1.645, 1.96, and 2.58.

Example 18. A physical therapist wished to estimate, with 99% confidence, the mean maximal strength of a particular muscle in a certain group of individuals. He is willing to assume that strength scores are approximately normally distributed with a variance of 144. A sample of 15 subjects who participated in the experiment yielded a mean of 84.3.

Solution: Our reliability coefficient is 2.58. The standard error is $\sigma_{\bar{x}} = 12/\sqrt{15} = 3.0984$. Our 99% confidence interval for μ is

$$84.3 \pm 2.58(3.0984)$$

$$84.3 \pm 8.0$$

$$[76.3, 92.3]$$

We say we are 99% confident that the population mean is between 76.3 and 92.3 since, in repeated sampling, 99% of all intervals that could be constructed in the manner just described would include the population mean.

Situations in which the variable of interest is approximately normally distributed with a known variance are so rare as to be almost nonexistent. The purpose of the preceding examples, which assumed that these ideal conditions existed, was to establish the theoretical background for constructing confidence intervals for population means. In most practical situations either the variables are not approximately normally distributed or the population variances are not known or both. The next example and the next section explain the procedures that are available for use in the less than ideal, but more common, situations.

Sampling from nonnormal populations As noted, it will not always be possible or prudent to assume that the population of interest is normally distributed. Thanks to the central limit theorem, this will not deter us if we are able to select a large enough sample. We have learned that for large samples, the sampling distribution of \bar{x} is approximately normally distributed regardless of how the parent population is distributed.

Example 19. Punctuality of patients in keeping appointments is of interest to a research team. In a study of patient flow through the offices of general practitioners, it was found that a sample of 35 patients were 17.2 minutes late for appointments, on the average. Previous research had shown the standard deviation to be about 8 minutes.

The population distribution was felt to be nonnormal. What is the 90% confidence interval for μ , the true mean amount of time late for appointments?

Solution: Since the sample size is fairly large ($n = 30$), and since the population standard deviation is known, we draw on the central limit theorem and assume that the sampling distribution of \bar{x} to be approximately normally distributed. For a 90% confidence interval, we know the reliability coefficient is 1.645. The standard error is $\sigma_{\bar{x}} = 8/\sqrt{35} = 1.3522$, so that our 90% confidence interval for μ is

$$17.2 \pm 1.645(1.3522)$$

$$17.2 \pm 2.2$$

$$[15.0, 19.4]$$

Frequently, when the sample is large enough for the application of the central limit theorem, the population variance is unknown. In that case, we use the sample variance as a replacement for the unknown population variance in the formula for constructing a confidence interval for the population mean.

3.3.3 t distribution

In our previous examples of confidence intervals, we assumed that the population variance is known. Practically speaking, this assumption is not very realistic. In most instances when a statistical study is being conducted, it is being done for the first time; there is no way to know prior to the study either the mean or the variance of the population of interest. Next, we consider the more realistic problem of constructing a confidence interval on a population mean when the population variance is assumed to be **unknown**.

To derive a general formula for a $100(1 - \alpha)\%$ confidence interval on μ under these circumstances, it is natural to begin by considering the random variable used earlier, namely

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

There are two problems to overcome:

1. The value of σ is not known and must be estimated;
2. The distribution of the random variable obtained by replacing σ by an estimator is not known.

The first problem is easy to overcome. We shall use the sample standard deviation S as an estimator for σ . The sample standard deviation is

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

The second problem is a little more difficult to solve. When we replace σ by its estimator S , the random variable $(\bar{X} - \mu)/(S\sqrt{n})$ results. When the sample size is large, say, greater than 30, our faith in s as an approximation of σ is usually substantial, and we may be appropriately justified in using normal distribution theory to construct a confidence interval for the population mean. In that event, we proceed as instructed in section (3.3.2).

It is when we have small samples that it becomes mandatory for us to find an alternative procedure for constructing confidence intervals.

It can be shown that the distribution of this random variable is no longer standard normal. Rather, when sampling from a normal distribution, it follows what is called a *Student's t*, or simply a t distribution. We pause briefly to consider this distribution.

Definition 15 (t distribution). Let Z be a standard normal random variable and let χ_γ^2 be an independent chi-squared random variable with γ degrees of freedom. The random variable

$$t = \frac{Z}{\sqrt{\chi_\gamma^2/\gamma}}$$

is said to follow a t distribution with γ degrees of freedom.

This definition implies that to show that a random variable follows a t distribution, we must show that it can be written as a ratio of a standard normal random variable to the square root of an independent chi-squared random variable divided by its degrees of freedom.

We note here the characteristics of t distribution that will be useful in the work that follows:

1. There are infinitely many t distribution, each identified by one parameter γ , called **degrees of freedom**. This parameter is always a positive integer. The notation t_γ denotes a t random variable with γ degrees of freedom.
2. Each t random variable is continuous. The density for a t random variable with γ degrees of freedom is given by

$$f(t) = \frac{\Gamma(\gamma + 1)/2}{\Gamma(\gamma/2)\sqrt{\pi\gamma}} \left(1 + \frac{t^2}{\gamma}\right)^{-(\gamma+1)/2} \quad -\infty < t < \infty$$

3. The graph of the density of a t_γ random variable is a symmetric bell-shaped curve centered at 0.
4. The parameter γ is a shape parameter in the sense that as its value increases, the variance of the random variable t_γ decreases. Thus as the value of γ increases, the bell-shaped curve associated with t_γ becomes more compact.
5. As the number of degrees of freedom increases, the bell-shaped curve associated with the t_γ random variable approaches the standard normal curve.
6. Compared to the normal distribution, the t distribution is less peaked in the center and has thicker tails. Figure 5 compares the t distribution with the normal.

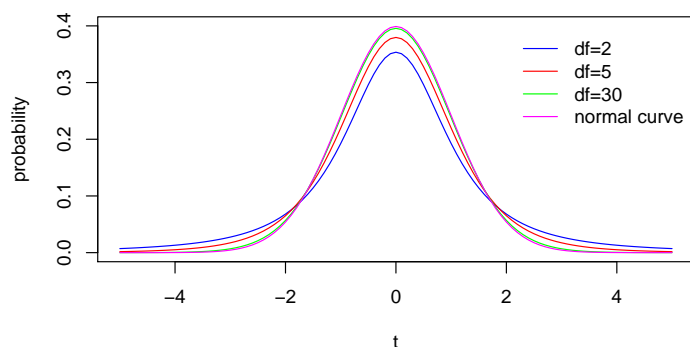


Figure 5: The t distribution with different degrees of freedom and a comparison with the normal distribution.

Theorem 7 (Distribution of $(n-1)S^2/\sigma^2$). Let X_1, X_2, \dots, X_n be a random sample from a normal distribution with mean μ and variance σ^2 . The random variable

$$(n-1)S^2/\sigma^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$

has a chi-squared distribution with $n-1$ degrees of freedom.

Theorem 8. Let X_1, X_2, \dots, X_n be a random sample from a normal distribution with mean μ and variance σ^2 . The random variable

$$\frac{\bar{X} - \mu}{S\sqrt{n}}$$

follows a t distribution with $n - 1$ degrees of freedom.

Proof. We shall show that the random variable $(\bar{X} - \mu)/(S\sqrt{n})$ can be written as the ratio of a standard normal random variable to the square root of an independent chi-squared random variable divided by its degrees of freedom. We know from Theorem 4 that \bar{X} is normal with mean μ and variance σ^2/n . Standardizing, $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is standard normal. By Theorem 7, $(n - 1)S^2/\sigma^2$ is a chi-squared random variable with $n - 1$ degrees of freedom. Consider the random variable

$$\frac{Z}{\sqrt{\chi_\gamma^2/\gamma}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(n - 1)S^2/\sigma^2(n - 1)}} = \frac{\bar{X} - \mu}{S\sqrt{n}}$$

Since \bar{X} and S are independent, this random variable follows a t distribution with $n - 1$ degrees of freedom. \square

Confidence interval using t The general procedure for constructing confidence interval is not affected by our having to use the t distribution rather than the standard normal distribution. We will make use of the relationship expressed by

$$\text{estimator} \pm (\text{reliability coefficient}) \times (\text{standard error of the estimator})$$

What is different is the source of the reliability coefficient. It is now obtained from the t distribution rather than from the standard normal distribution.

It is now easy to determine the general form for a $100(1 - \alpha)\%$ confidence interval on μ when σ^2 is unknown. We need only note that the two random variables

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad t_\gamma = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

have the same algebraic structure.

Theorem 9 ($100(1 - \alpha)\%$ confidence interval on μ when σ^2 is unknown). Let X_1, X_2, \dots, X_n be a random sample from a normal distribution with mean μ and variance σ^2 . A $100(1 - \alpha)\%$ confidence interval on μ is given by

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} \tag{3}$$

We emphasize that a requirement for the strictly valid use of the t distribution is that the sample must be drawn from a normal distribution. Experience has shown, however, that moderate departures from this requirement can be tolerated. As a consequence, the t distribution is used even when it is known that the parent population deviated somewhat from normality. Most researchers require that an assumption of, at least, a mound-shaped population distribution be tenable.

Example 20. The effectiveness of early weightbearing and ankle mobilization therapies following acute repair of a ruptured Achilles tendon. One of the variables they measured following treatment was the isometric gastrocsoleus muscle strength. In 19 subjects, the mean isometric strength for the operated limb was 250.8 (in newtons) with a standard deviation of 130.9. We assume that these 19 patients constitute a random sample from a population of similar subjects. We wish to use these sample data to estimate for the population the mean isometric strength after surgery.

Solution: We may use the sample mean, 250.8, as a point estimate of the population mean, but, because the population standard deviation is unknown, we must assume the population of values to be at least approximately normally distributed before constructing a confidence interval for μ .

Let us assume that such an assumption is reasonable and that a 95% confidence interval is desired. We have our estimator, \bar{x} , and our standard error is $s/\sqrt{n} = 130.9/\sqrt{19} = 30.0305$. We need now to find the reliability coefficient, the value of t associated with a confidence coefficient of .95 and $n - 1 = 18$ degrees of freedom. Since a 95% confidence interval leaves .05 of the area under the curve of t to be equally divided between the two tails, we need the value of t to the right of which lies .025 of the area. This is the value of t to the left of which lies .975 of the area under the curve. This t can be obtained using the following R command

```
qt(0.975,df=18)
```

and we get 2.1009. The area to the right of this value is equal to the desired .025. We now construct our 95% confidence interval as follows:

$$250.8 \pm 2.1009(30.0305)$$

$$250.8 \pm 63.1$$

$$[187.7, 313.9]$$

This interval may be interpreted from both the probabilistic and practical points of view. We are 95% confident that the true population mean, μ , is somewhere between 177.7 and 313.9 because, in repeated sampling, 95% of intervals constructed in like manner will include μ .

Example 21. Compute a 95% confidence interval based on the following sample of size 10:

97, 117, 140, 78, 99, 148, 108, 135, 126, 121

Solution: We have

$$n = 10, \bar{x} = 116.90, s = 21.70$$

Because we want a 95% confidence interval, $\alpha = .05$. Therefore, from Equation (3), the 95% confidence interval is

$$[116.9 - t_{0.025}21.7/\sqrt{10}, 116.9 + t_{0.025}21.7/\sqrt{10}]$$

Since $t_{0.025} = 2.262$ (R command: `qt(0.975, df=9)`), the interval is

$$[116.9 - (2.262)(21.7/\sqrt{10}), 116.9 + (2.262)(21.7/\sqrt{10})]$$

Deciding between z and t When we construct a confidence interval for a population mean, we must decide whether to use a value of z or a value of t as the reliability factor. To make an appropriate choice we must consider sample size, whether the sampled population is normally distributed, and whether the population variance is known. Figure provides a flowchart that one can use to decide quickly whether the reliability factor should be z or t (add the figure here).

3.4 Estimation of the variance of a distribution

3.4.1 Point estimation

We have defined the sample variance S^2 as in Eq. (1). This was done so that the resulting estimator would be unbiased for σ^2 . We now prove that this is the case.

Theorem 10. Let S^2 be the sample variance based on a random sample of size n from a distribution with mean μ and variance σ^2 . S^2 is an unbiased estimator for σ^2 .

Proof. By definition,

$$\begin{aligned}
E[S^2] &= E \left[\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1} \right] \\
&= \frac{1}{n-1} E \left[\sum_{i=1}^n (X_i - \mu + \mu - \bar{X})^2 \right] \\
&= \frac{1}{n-1} E \left[\sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + \sum_{i=1}^n (\bar{X} - \mu)^2 \right] \\
&= \frac{1}{n-1} E \left[\sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \frac{n(\sum X_i - n\mu)}{n} + n(\bar{X} - \mu)^2 \right] \\
&= \frac{1}{n-1} E \left[\sum_{i=1}^n (X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2 \right] \\
&= \frac{1}{n-1} E \left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right] \\
&= \frac{1}{n-1} \left[\sum_{i=1}^n E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2] \right]
\end{aligned}$$

Note that since X_1, X_2, \dots, X_n is a random sample from a distribution with variance σ^2 , $E[(X_i - \mu)^2] = \sigma^2$ for each $i = 1, 2, \dots, n$. Note that by Theorem 3, $Var[\bar{X}] = E[(\bar{X} - \mu)^2] = \sigma^2/n$. By substitution, we obtain

$$E[S^2] = \frac{1}{n-1} \left[\sum_{i=1}^n \sigma^2 - n\sigma^2/n \right] = \frac{1}{n-1} (n\sigma^2 - \sigma^2) = \sigma^2$$

□

It should be noted that even though S^2 is an unbiased estimator for σ^2 , it can be shown that S is not unbiased for σ . This emphasizes the fact that unbiasedness is desirable in an estimator, but not essential.

The standard error of the mean is given by σ/\sqrt{n} and is estimated by s/\sqrt{n} . This statistic is called the **sample standard error** and will be used extensively later.

Example 22. Suppose that a woman wants to estimate her exact day of ovulation for conceptive / contraceptive purposes. A theory exists that at the time of ovulation the body temperature rises by an amount from 0.5 to 1.0 degrees Fahrenheit. Thus changes in body temperature can be used to guess the day of ovulation.

To use this method, we need a good estimate of basal body temperature during a period when ovulation is definitely not occurring. Suppose that a woman measures her body temperature on awakening on the first 10 days after menstruation and obtains the following data:

97.2, 96.8, 97.4, 97.4, 97.3, 97.0, 97.1, 97.3, 97.2, 97.3

What is the best estimate of her underlying basal body temperature μ ? How precise is this estimate? Estimate the variance of the distribution of basal body temperature.

Solution: The best estimate of her underlying body temperature during the nonovulation period is given by

$$\begin{aligned}\bar{x} &= \frac{97.2 + 96.8 + 97.4 + 97.4 + 97.3 + 97.0 + 97.1 + 97.3 + 97.2 + 97.3}{10} \\ &= 97.2\end{aligned}$$

An unbiased estimate of the variance of the body temperature is

$$s^2 = \frac{n \sum_{i=1}^{10} x_i^2 - (\sum_{i=1}^{10} x_i)^2}{n(n-1)} = 0.0355$$

The standard error of the estimate for μ is given by

$$\frac{s}{\sqrt{10}} = 0.189/\sqrt{10} = 0.06$$

The intuitive estimator for σ^2 with n in the denominator rather than $n-1$, that is,

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

tends to underestimate the underlying variance σ^2 by a factor of $(n-1)/n$. This factor is considerable for small samples but tends to be negligible for large samples.

3.4.2 Interval estimation

The problem of interval estimation of the mean of a normal distribution has been discussed. We often want to obtain interval estimates of the variance as well. As was the case for the mean, the interval estimates will hold exactly only if the underlying distribution is normal. The interval estimates perform much more poorly for the variance than for the mean if the underlying distribution is not normal, and they should be used with caution in this case.

Example 23. An arteriosonde machine “prints” blood-pressure readings on a tape so that the measurement can be read rather than heard. A major argument for using such a machine is that the variability of measurements obtained by different observers on the same person will be lower than with a standard blood-pressure cuff.

Suppose that we have the data in Table 2, consisting of systolic blood-pressure measurements obtained on 10 people and read by 2 observers. We use the difference d_i between the first and second observer to assess interobserver variability. In particular, if we assume the underlying distribution of these differences is normal with mean μ and variance σ^2 , then it is of primary interest to estimate σ^2 . The higher σ^2 is, the higher the interobserver variability.

Table 2: Systolic blood pressure measurements (mm Hg) from an arteriosonde machine obtained from 10 people and read by 2 observers

Person (i)	1	2	difference d_i
1	194	200	-6
2	126	123	+3
3	130	128	+2
4	98	101	-3
5	136	135	+1
6	145	145	0
7	110	111	-1
8	108	107	+1
9	102	99	+3
10	126	128	-2

$$\text{mean difference} = (-6 + 3 + \dots - 2)/10 = -0.2 = \bar{d}$$

$$\text{sample variance} = s^2 = \frac{1}{9} \sum_{i=1}^{10} (d_i - \bar{d})^2 = 8.178$$

How can an interval estimate for σ^2 be obtained?

To obtain an interval estimate of σ^2 , we need to find the sampling distribution of S^2 . From Theorem 7, we know that $(n - 1)S^2/\sigma^2$ has a chi-squared distribution with $n - 1$ degrees of freedom. To obtain a $100(1 - \alpha)\%$ confidence interval for σ^2 , we first obtain the $100(1 - \alpha)\%$ confidence interval for $(n - 1)s^2/\sigma^2$. To do this, we select the values of χ^2 in such a way that $\alpha/2$ is to the left of the smaller value and $\alpha/2$ is to the

right of the larger value. In other words, the two values of χ^2 are selected in such a way that α is divided equally between the two tails of the distribution. We may designate these two values of χ^2 as $\chi_{\alpha/2}^2$ and $\chi_{1-(\alpha/2)}^2$, respectively. The $100(1 - \alpha)\%$ confidence interval for $(n - 1)S^2/\sigma^2$ then is given by

$$\chi_{\alpha/2}^2 < \frac{(n - 1)s^2}{\sigma^2} < \chi_{1-(\alpha/2)}^2$$

We now manipulate this expression in such a way that we obtain an expression with σ^2 alone as the middle term. First, let us divide each term by $(n - 1)s^2$ to get

$$\frac{\chi_{\alpha/2}^2}{(n - 1)s^2} < \frac{1}{\sigma^2} < \frac{\chi_{1-(\alpha/2)}^2}{(n - 1)s^2}$$

If we take the reciprocal of this expression, we have

$$\frac{(n - 1)s^2}{\chi_{1-(\alpha/2)}^2} < \sigma^2 < \frac{(n - 1)s^2}{\chi_{\alpha/2}^2}$$

which is the $100(1 - \alpha)\%$ confidence interval for σ^2 . If we take the square root of each term, we have the following $100(1 - \alpha)\%$ confidence interval for σ , the population standard deviation:

$$\sqrt{\frac{(n - 1)s^2}{\chi_{1-(\alpha/2)}^2}} < \sigma < \sqrt{\frac{(n - 1)s^2}{\chi_{\alpha/2}^2}}$$

Example 24. In a study of the effectiveness of a gluten-free diet in first-degree relatives of patients with type I diabetics, seven subjects were placed on a gluten-free diet for 12 months. Prior to the diet, baseline measurements of several antibodies and autoantibodies were taken, one of which was the diabetes related insulin autoantibody (IAA). The IAA levels were measured by radiobinding assay. The seven subjects had IAA units of

9.7, 12.3, 11.2, 5.1, 14.8, 17.7

We wish to estimate from the data in this sample the variance of the IAA units in the population from which the sample was drawn and construct a 95% confidence interval for this estimate.

Solution: The sample yielded a value of $s^2 = 39.763$. The degrees of freedom are $n - 1 = 6$. The appropriate values of χ^2 are

$$\chi_{1-(\alpha/2)}^2 = 14.449, \quad \chi_{\alpha/2}^2 = 1.237$$

These values can be obtained using R command `qchisq(.975, df=6)` and `qchisq(.025, df=6)`. Our 95% confidence interval for σ^2 is

$$\frac{6(39.763)}{14.449} < \sigma^2 < \frac{6(39.763)}{1.237} \Rightarrow 16.512 < \sigma^2 < 192.868$$

The 95% confidence interval for σ is

$$4.063 < \sigma < 13.888$$

We are 95% confident that the parameters being estimated are within the specified limits, because we know that in the long run, in repeated sampling, 95% of intervals constructed as illustrated would include the respective parameters.

Precautions: Although this method of constructing confidence intervals for σ^2 is widely used, it is not without its drawbacks. First, the assumption of the normality of the population from which the sample is drawn is crucial, and results may be misleading if the assumption is ignored.

Although difficulty with these intervals results from the fact that the estimator is not in the center of the confidence interval, as is the case with the confidence interval for μ . This is because the chi-square distribution, unlike the normal, is not symmetric. The practical implication of this is that the method for the construction of confidence intervals for σ^2 , does not yield the shortest possible confidence intervals.

3.5 Estimation of the difference between two sample means

Frequently the interest in an investigation is focused on two populations. Specifically, an investigator may wish to know something about the difference between two population means. In one investigation, for example, a researcher may wish to know if it is reasonable to conclude that two population means are different. In another situation, the researcher may desire knowledge about the magnitude of the difference between two population means. If the researchers are able to conclude that the population means are different, they may wish to know by how much they differ. A knowledge of the sampling distribution of the difference between two means is useful in investigations of this type.

Example 25. Suppose we have two populations of individuals. Population 1 has experienced some condition thought to be associated with mental retardation, and population

2 has not experienced the condition. The distribution of intelligence scores in each of the two populations is believed to be approximately normally distributed with a standard deviation of 20.

Suppose, further, that we take a sample of 15 individuals from each population and compute for each sample the mean intelligence score with following results: $\bar{x}_1 = 92$ and $\bar{x}_2 = 105$. If there is no difference between the two populations, with respect to their true mean intelligence scores, what is the probability of observing a difference this large or larger between two sample means?

Solution: To answer this question, we need to know the nature of the sampling distribution of the relevant statistic, the **difference between two sample means**, $\bar{x}_1 - \bar{x}_2$. Notice we seek a probability associated with the difference between two sample means rather than a single mean.

3.5.1 Construction of the sampling distribution of $\bar{x}_1 - \bar{x}_2$

Although, in practice, we would not attempt to construct the desired sampling distribution, we can conceptualize the manner in which it could be done when sampling is from finite populations. We would begin by selecting from population 1 all possible samples of size 15 and computing the mean for each sample. We know that there would be $\binom{N_1}{n_1}$ such samples where N_1 is the population size and $n_1 = 15$. Similarly, we would select all possible samples of size 15 from population 2 and compute the mean for each of these samples. We would then take all possible pairs of sample means, one from population 1 and one from population 2, and take the difference.

3.5.2 Characteristics of the sampling distribution of $\bar{x}_1 - \bar{x}_2$

Theorem 11. Given two normally distributed populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , respectively, the sampling distribution of the difference, $\bar{x}_1 - \bar{x}_2$, between the means of independent samples of size n_1 and n_2 drawn from these populations:

1. is normally distributed
2. has mean $\mu_1 - \mu_2$
3. has variance $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Sampling from nonnormal populations Many times a research is faced with one or the other of the following problems: the necessity of (1) sampling from nonnormally

distributed populations, or (2) sampling from populations whose functional forms are not known. A solution to these problems is to take large samples, since when the sample sizes are large the central limit theorem applies and the distribution of the difference between two sample means is at least approximately normally distributed with a mean equal to $\mu_1 - \mu_2$ and a variance of $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$. To find probabilities associated with specific values of the statistic, the, our procedure would be the same as that given when sampling is from normally distributed populations.

Converting to z The normal distribution in theorem (11) can be transformed to the standard normal distribution by:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Example 26. Suppose that it has been established that for a certain type of client the average length of a home visit by a public health nurse is 45 minutes with a standard deviation of 15 minutes, and that for a second type of client the average home visit is 30 minutes long with a standard deviation of 20 minutes. If a nurse randomly visits 35 clients from the first and 40 from the second population, what is the probability that the average length of home visit will differ between the two groups by 20 or more minutes?

Solution: No mention is made of the functional form of the two populations, so let us assume that this characteristic is unknown, or that the populations are not normally distributed. Since the sample sizes are large (greater than 30) in both cases, we draw on the results of the central limit theorem to answer the question posed. We know that the difference between sample means is at least approximately normally distributed with the following mean and variance:

$$\begin{aligned}\mu_{\bar{x}_1 - \bar{x}_2} &= \mu_1 - \mu_2 = 45 - 30 = 15 \\ \sigma_{\bar{x}_1 - \bar{x}_2}^2 &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{15^2}{35} + \frac{20^2}{40} = 16.4286\end{aligned}$$

The area under the curve of $\bar{x}_1 - \bar{x}_2$ that we seek is that area to the right of 20. The corresponding value of z in the standard normal is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{20 - 15}{\sqrt{16.4286}} = 1.23$$

We can find that the area to the right of $z = 1.23$ is .1093.

3.5.3 Confidence interval

Sampling from normal distributions When the population variances are known and the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal, the $100(1 - \alpha)$ percent confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{(1-\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

An examination of a confidence interval for the difference between population means provides information that is helpful in deciding whether or not it is likely that the two population means are equal. When the constructed interval does not include zero, we say that the interval provides evidence that the two population means are not equal. When the interval includes zero, we say that the population means may be equal.

Example 27. A research team is interested in the difference between serum uric acid levels in patients with and without Down's syndrome. In a large hospital for the treatment of the mentally retarded, a sample of 12 individuals with Down's Syndrome yielded a mean of $\bar{x} = 4.5$ mg/100 ml. In a general hospital a sample of 15 normal individuals of the same age and sex were found to have a mean value of $\bar{x}_2 = 3.4$. If it is reasonable

to assume that the two populations of values are normally distributed with variances equal to 1 and 1.5, find the 95% confidence interval for $\mu_1 - \mu_2$.

Solution: For a point estimate of $\mu_1 - \mu_2$, we use $\bar{x}_1 - \bar{x}_2 = 4.5 - 3.4 = 1.1$. The reliability coefficient corresponding to .95 is 1.96. The standard error is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{1}{12} + \frac{1.5}{15}} = .4282$$

The 95% confidence interval, then, is

$$\begin{aligned} &1.1 \pm 1.96(.4282) \\ &1.1 \pm .84 \\ &[.26, 1.94] \end{aligned}$$

We say that we are 95% confident that the true difference $\mu_1 - \mu_2$, is somewhere between .26 and 1.94 because, in repeated sampling, 95% of the intervals constructed in this manner would include the difference between the true means. Since the interval does not include zero, we conclude that the two population means are not equal.

Sampling from nonnormal populations The construction of a confidence interval for the difference between two populations means when sampling is from nonnormal populations proceeds in the same manner as in Example 27 if the sample sizes n_1 and n_2 are large. Again, this is a result of the central limit theorem. If the population variances are unknown, we use the sample variances to estimate them.

Example 28. Despite common knowledge of the adverse effects of doing so, many women continue to smoke while pregnant. The effectiveness of a smoking cessation program for pregnant women is examined. The mean number of cigarettes smoked daily at the close of the program by the 328 women who completed the program was 4.3 with a standard deviation of 5.22. Among 64 women who did not complete the program, the mean number of cigarettes smoked per day at the close of the program was 13 with a standard deviation of 8.97. We wish to construct a 99% confidence interval for the difference between the means of the populations from which the samples may be presumed to have been selected.

Solution: No information is given regarding the shape of the distribution of the distribution of cigarettes smoked per day. Since our sample sizes are large, however, the central limit theorems assures us that the sampling distribution of the difference between sample means will be approximately normally distributed even if the distribution of the variable in the populations is not normally distributed. We may use the fact as justification for using the z statistic as the reliability factor in the construction of our confidence interval. Also, since the population standard deviations are not given, we will use the sample standard deviations to estimate them. The point estimate for the difference between population means is the difference between sample means, $4.3-13.0=-8.7$. For 99% confidence interval, the reliability factor is 2.58. The estimated standard error is

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{5.22^2}{328} + \frac{8.97^2}{64}} = 1.1577$$

The 99% confidence interval for the difference between population means is

$$\begin{aligned} & -8.7 \pm 2.58(1.1577) \\ & [-11.7, -5.7] \end{aligned}$$

We are 99% confident that the mean number of cigarettes smoked per day for women who complete the program is between 5.7 and 11.7 lower than for women who do not complete the program.

References

- [1] J. S. Milton and Jesse C. Arnold. *Introduction to probability and statistics: principles and applications for engineering and the computing sciences*. McGraw-Hill, Inc., 1995.
- [2] Bernard Rosner. *Fundamentals of biostatistics*. Thomson Brooks/Cole, 2006.