Mathematics for Data Science

Lecture 4: Confidence Intervals

Mohamad GHASSANY

EFREI Paris

INTRODUCTION

Engineers are often involved in estimating parameters. For example, there is an ASTM Standard E23 that defines a technique called the Charpy V-notch method for notched bar impact testing of metallic materials. The impact energy is often used to determine if the material experiences a ductile-to-brittle transition as the temperature decreases. Suppose that you have tested a sample of 10 specimens of a particular material with this procedure. You know that you can use the sample average $\overline{X}$ to estimate the true mean impact energy μ. However, we also know that the true mean impact energy is unlikely to be exactly equal to your estimate. Reporting the results of your test as a single number is unappealing, because there is nothing inherent in $\overline{X}$ that provides any information about how close it is to μ. Your estimate could be very close, or it could be considerably far from the true mean. A way to avoid this is to report the estimate in terms of a range of plausible values called a confidence interval. A confidence interval always specifies a confidence level, usually 90%, 95%, or 99%, which is a measure of the reliability of the procedure. So if a 95% confidence interval on the impact energy based on the data from your 10 specimens has a lower limit of 63.84J and an upper limit of 65.08J, then we can say that at the 95% level of confidence any value of mean impact energy between 63.84 J and 65.08 J is a plausible value. By reliability, we mean that if we repeated this experiment over and over again, 95% of all samples would produce a confidence interval that contains the true mean impact energy, and only 5% of the time would the interval be in error. In this chapter you will learn how to construct confidence intervals and other useful types of statistical intervals for many important types of problem situations.

In the previous chapter, we illustrated how a point estimate of a parameter can be estimated from sample data. However, it is important to understand how good the estimate obtained is. For example, suppose that we estimate the mean viscosity of a chemical product to be $\hat{\mu} = \overline{x} = 1000$. Now because of sampling variability, it is almost never the case that the true mean μ is exactly equal to the estimate $\overline{x}$. The point estimate says nothing about how close $\hat{\mu}$ is to μ. Is the process mean likely to be between 900 and 1100? Or is it likely to be between 990 and 1010? The answer to these questions affects our decisions regarding this process. Bounds that represent an interval of plausible values for a parameter are examples of an interval estimate. Surprisingly, it is easy to determine such intervals in many cases, and the same data that provided the point estimate are typically used.

An interval estimate for a population parameter is called a **confidence interval**. Information about the precision of estimation is conveyed by the length of the interval. A short interval implies precise estimation. We cannot be certain that the interval contains the true, unknown population parameter—we use only a sample from the full population to compute the point estimate and the interval. However, the confidence interval is constructed so that we have high confidence that it does contain the unknown population parameter. Confidence intervals are widely used in engineering and the sciences.

8-1 Confidence Interval on the Mean of a Normal Distribution, Variance Known

The basic ideas of a confidence interval (CI) are most easily understood by initially considering a simple situation. Suppose that we have a normal population with unknown mean μ and known variance σ^2. This is a somewhat unrealistic scenario because typically both the mean and variance are unknown. However, in subsequent sections, we will present confidence intervals for more general situations.

8-1.1 DEVELOPMENT OF THE CONFIDENCE INTERVAL AND ITS BASIC PROPERTIES

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with unknown mean μ and known variance σ^2. From the results of Chapter 5, we know that the sample mean $\bar{X}$ is normally distributed with mean μ and variance σ^2/n. We may **standardize** $\bar{X}$ by subtracting the mean and dividing by the standard deviation, which results in the variable

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \tag{8-1}$$

The random variable Z has a standard normal distribution.

A **confidence interval** estimate for μ is an interval of the form $l \le \mu \le u$, where the end-points l and u are computed from the sample data. Because different samples will produce different values of l and u, these end-points are values of random variables L and U, respectively. Suppose that we can determine values of L and U such that the following probability statement is true:

$$P\{L \le \mu \le U\} = 1 - \alpha \tag{8-2}$$

where $0 \le \alpha \le 1$. There is a probability of $1 - \alpha$ of selecting a sample for which the CI will contain the true value of μ. Once we have selected the sample, so that $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, and computed l and u, the resulting **confidence interval** for μ is

$$l \le \mu \le u \tag{8-3}$$

The end-points or bounds l and u are called the **lower- and upper-confidence limits (bounds)**, respectively, and $1 - \alpha$ is called the **confidence coefficient**.

In our problem situation, because $Z = (\bar{X} - \mu)/(\sigma / \sqrt{n})$ has a standard normal distribution, we may write

$$P\left\{-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le z_{\alpha/2}\right\} = 1 - \alpha$$

Now manipulate the quantities inside the brackets by (1) multiplying through by $\sigma/\sqrt{n}$, (2) subtracting $\bar{X}$ from each term, and (3) multiplying through by −1. This results in

$$P\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha \quad (8\text{-}4)$$

This is a **random interval** because the end-points $\bar{X} \pm Z_{\alpha/2}\sigma/\sqrt{n}$ involve the random variable $\bar{X}$. From consideration of Equation 8-4, the lower and upper end-points or limits of the inequalities in Equation 8-4 are the lower- and upper-confidence limits L and U, respectively. This leads to the following definition.

Confidence Interval on the Mean, Variance Known

If $\bar{x}$ is the sample mean of a random sample of size n from a normal population with known variance σ^2, a $100(1-\alpha)\%$ CI on μ is given by

$$\bar{x} - z_{\alpha/2}\,\sigma/\sqrt{n} \le \mu \le \bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n} \quad (8\text{-}5)$$

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution.

The development of this CI assumed that we are sampling from a normal population. The CI is quite robust to this assumption. That is, moderate departures from normality are of no serious concern. From a practical viewpoint, this implies that an **advertised** 95% CI might have actual confidence of 93% or 94%.

Example 8-1 **Metallic Material Transition** ASTM Standard E23 defines standard test methods for notched bar impact testing of metallic materials. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (J) on specimens of A238 steel cut at 60ºC are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, and 64.3. Assume that impact energy is normally distributed with $\sigma = 1J$. We want to find a 95% CI for μ, the mean impact energy. The required quantities are $z_{\alpha/2} = z_{0.025} = 1.96$, $n = 10$, $\sigma = 1$, and $\bar{x} = 64.46$. The resulting 95% CI is found from Equation 8-5 as follows:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$64.46 - 1.96\frac{1}{\sqrt{10}} \leq \mu \leq 64.46 + 1.96\frac{1}{\sqrt{10}}$$

$$63.84 \leq \mu \leq 65.08$$

Practical Interpretation: Based on the sample data, a range of highly plausible values for mean impact energy for A238 steel at 60°C is $63.84J \leq \mu \leq 65.08J$.

Interpreting a Confidence Interval

How does one interpret a confidence interval? In the impact energy estimation problem in Example 8-1, the 95% CI is $63.84 \leq \mu \leq 65.08$, so it is tempting to conclude that μ is within this interval with probability 0.95. However, with a little reflection, it is easy to see that this cannot be correct; the true value of μ is unknown, and the statement $63.84 \leq \mu \leq 65.08$ is either correct (true with probability 1) or incorrect (false with probability 1). The correct interpretation lies in the realization that a CI is a *random interval* because in the probability statement defining the end-points of the interval (Equation 8-2), L and U are random variables. Consequently, the correct interpretation of a $100(1-\alpha)\%$ CI depends on the relative frequency view of probability. Specifically, if an infinite number of random samples are collected and a $100(1-\alpha)\%$ confidence interval for μ is computed from each sample, $100(1-\alpha)\%$ of these intervals will contain the true value of μ.

The situation is illustrated in Fig. 8-1, which shows several $100(1-\alpha)\%$ confidence intervals for the mean μ of a normal distribution. The dots at the center of the intervals indicate the point estimate of μ (that is, $\bar{x}$). Notice that one of the intervals fails to contain the true value of μ. If this were a 95% confidence interval, in the long run only 5% of the intervals would fail to contain μ.

Now in practice, we obtain only one random sample and calculate one confidence interval. Because this interval either will or will not contain the true value of μ, it is not reasonable to attach a probability level to this specific event. The appropriate statement is that the observed interval $[l, u]$ brackets the true value of μ with **confidence** $100(1-\alpha)$. This statement has a frequency interpretation; that is, we do not know whether the statement is true for this specific sample, but the *method* used to obtain the interval $[l, u]$ yields correct statements $100(1-\alpha)\%$ of the time.

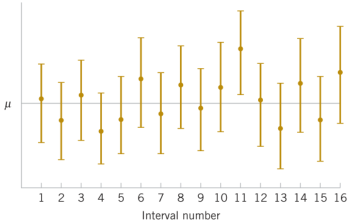

FIGURE 8-1 Repeated construction of a confidence interval for μ.

Confidence Level and Precision of Estimation

Notice that in Example 8-1, our choice of the 95% level of confidence was essentially arbitrary. What would have happened if we had chosen a higher level of confidence, say, 99%? In fact, is it not reasonable that we would want the higher level of confidence? At $\alpha = 0.01$, we find $z_{\alpha/2} = z_{0.01/2} = z_{0.005} = 2.58$, while for $\alpha = 0.05$, $z_{0.025} = 1.96$. Thus, the **length** of the 95% confidence interval is

$$2\left(1.96\sigma/\sqrt{n}\right) = 3.92\sigma/\sqrt{n}$$

whereas the length of the 99% CI is

$$2\left(2.58\sigma/\sqrt{n}\right) = 5.16\sigma/\sqrt{n}$$

Thus, the 99% CI is longer than the 95% CI. This is why we have a higher level of confidence in the 99% confidence interval. Generally, for a fixed sample size n and standard deviation σ, the higher the confidence level, the longer the resulting CI.

The length of a confidence interval is a measure of the **precision** of estimation. Many authors define the half-length of the CI (in our case $z_{\alpha/2}\sigma/\sqrt{n}$) as the bound on the error in estimation of the parameter. From the preceeding discussion, we see that precision is inversely related to the confidence level. It is desirable to obtain a confidence interval that is short enough for decision-making purposes and that also has adequate confidence. One way to achieve this is by choosing the sample size n to be large enough to give a CI of specified length or precision with prescribed confidence.

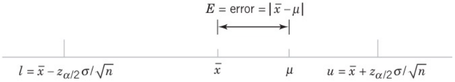

FIGURE 8-2 Error in estimating μ with $\bar{x}$.

8-1.3 ONE-SIDED CONFIDENCE BOUNDS

The confidence interval in Equation 8-5 gives both a lower confidence bound and an upper confidence bound for μ. Thus, it provides a two-sided CI. It is also possible to obtain one-sided confidence bounds for m by setting either the lower bound $l = -\infty$ or the upper bound $u = \infty$ and replacing $z_{\alpha/2}$ by z_α.

One-Sided Confidence Bounds on the Mean, Variance Known

A $100(1-\alpha)\%$ **upper-confidence bound** for μ is

$$\mu \le \bar{x} + z_\alpha \sigma / \sqrt{n} \quad (8\text{-}7)$$

and a $100(1-\alpha)\%$ **lower-confidence bound** for μ is

$$\bar{x} - z_\alpha \sigma / \sqrt{n} = l \le \mu \quad (8\text{-}8)$$

8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

When we are constructing confidence intervals on the mean μ of a normal population when σ^2 is known, we can use the procedure in Section 8-1.1. This CI is also approximately valid (because of the central limit theorem) regardless of whether or not the underlying population is normal so long as n is reasonably large ($n \geq 40$, say). As noted in Section 8-1.5, we can even handle the case of unknown variance for the large-sample-size situation. However, when the sample is small and σ^2 is unknown, we must make an assumption about the form of the underlying distribution to obtain a valid CI procedure. A reasonable assumption in many cases is that the underlying distribution is normal.

Many populations encountered in practice are well approximated by the normal distribution, so this assumption will lead to confidence interval procedures of wide applicability. In fact, moderate departure from normality will have little effect on validity. When the assumption is unreasonable, an alternative is to use nonparametric statistical procedures that are valid for any underlying distribution.

Suppose that the population of interest has a normal distribution with unknown mean μ and unknown variance σ^2. Assume that a random sample of size n, say, $X_1, X_2, \ldots, X_n$, is available, and let $\bar{X}$ and S^2 be the sample mean and variance, respectively.

We wish to construct a two-sided CI on μ. If the variance σ^2 is known, we know that $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ has a standard normal distribution. When σ^2 is unknown, a logical procedure is to replace σ with the sample standard deviation S. The random variable Z now becomes $T = (\bar{X} - \mu)/(S/\sqrt{n})$. A logical question is what effect replacing σ with S has on the distribution of the random variable T. If n is large, the answer to this question is "very little," and we can proceed to use the confidence interval based on the normal distribution from Section 8-1.5. However, n is usually small in most engineering problems, and in this situation, a different distribution must be employed to construct the CI.

***t* Distribution**

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

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{8-13}$$

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

The t probability density function is

$$f(x) = \frac{\Gamma[(k+1)/2]}{\sqrt{\pi k}\,\Gamma(k/2)} \cdot \frac{1}{[(x^2/k)+1]^{(k+1)/2}} \qquad -\infty < x < \infty \tag{8-14}$$

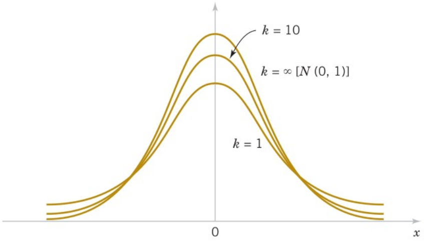

FIGURE 8-4 Probability density functions of several *t* distributions.

8-2.2 *t* CONFIDENCE INTERVAL ON μ

It is easy to find a 100(1 – α)% confidence interval on the mean of a normal distribution with unknown variance by proceeding essentially as we did in Section 8-1.1. We know that the distribution of $T = (\bar{X} - \mu)/(S/\sqrt{n})$ is t with $n - 1$ degrees of freedom. Letting $t_{\alpha/2,n-1}$ be the upper $100\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom, we may write

$$P\left(-t_{\alpha/2,n-1} \le T \le t_{\alpha/2,n-1}\right) = 1 - \alpha$$

or

$$P\left(-t_{\alpha/2,n-1} \le \frac{\bar{X} - \mu}{S/\sqrt{n}} \le t_{\alpha/2,n-1}\right) = 1 - \alpha$$

Rearranging this last equation yields

$$P\left(\bar{X} - t_{\alpha/2,n-1}\, S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,n-1}\, S/\sqrt{n}\right) = 1 - \alpha \qquad (8\text{-}15)$$

This leads to the following definition of the 100(1 – α)% two-sided confidence interval on μ.

Confidence Interval on the Mean, Variance Unknown

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a **100(1 – α)% confidence interval on μ** is given by

$$\bar{x} - t_{\alpha/2,n-1}\, s/\sqrt{n} \le \mu \le \bar{x} + t_{\alpha/2,n-1}\, s/\sqrt{n} \qquad (8\text{-}16)$$

where $t_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom.

The assumption underlying this CI is that we are sampling from a normal population. However, the t distribution-based CI is relatively insensitive or robust to this assumption. Checking the normality assumption by constructing a normal probability plot of the data is a good general practice. Small to moderate departures from normality are not a cause for concern.

One-sided confidence bounds on the mean of a normal distribution are also of interest and are easy to find. Simply use only the appropriate lower or upper confidence limit from Equation 8-16 and replace $t_{\alpha/2,n-1}$ by $t_{\alpha,n-1}$.

8-3 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Sometimes confidence intervals on the population variance or standard deviation are needed. When the population is modeled by a normal distribution, the tests and intervals described in this section are applicable. The following result provides the basis of constructing these confidence intervals.

χ^2 Distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2, and let S^2 be the sample variance. Then the random variable

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \qquad (8\text{-}17)$$

has a chi-square (χ^2) distribution with $n - 1$ degrees of freedom.

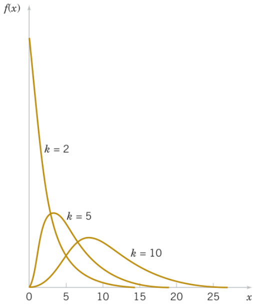

FIGURE 8-8 Probability density functions of several χ^2 distributions.

The probability density function of a χ^2 random variable is

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \qquad x > 0 \qquad (8\text{-}18)$$

where k is the number of degrees of freedom. The mean and variance of the χ^2 distribution are k and $2k$, respectively. Several chi-square distributions are shown in Fig. 8-8. Note that the

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

This leads to the following definition of the confidence interval for σ^2.

Confidence Interval on the Variance

If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2, then a **100(1 – α)% confidence interval on σ^2** is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \qquad \text{(8-19)}$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n - 1$ degrees of freedom, respectively. A confidence interval for σ has lower and upper limits that are the square roots of the corresponding limits in Equation 8-19.

It is also possible to find a 100(1 – α)% lower confidence bound or upper confidence bound on σ^2.

One-Sided Confidence Bounds on the Variance

The 100(1 – α)% lower and upper confidence bounds on σ^2 are

$$\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \le \sigma^2 \quad \text{and} \quad \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}} \qquad \text{(8-20)}$$

respectively.

The CIs given in Equations 8-19 and 8-20 are less robust to the normality assumption. The distribution of $(n-1)S^2/\sigma^2$ can be very different from the chi-square if the underlying population is not normal.

8-4 Large-Sample Confidence Interval for a Population Proportion

It is often necessary to construct confidence intervals on a population proportion. For example, suppose that a random sample of size n has been taken from a large (possibly infinite) population and that $X(\leq n)$ observations in this sample belong to a class of interest. Then $\hat{P} = X/n$ is a point estimator of the proportion of the population p that belongs to this class. Note that n and p are the parameters of a binomial distribution. Furthermore, from Chapter 4 we know that the sampling distribution of $\hat{P}$ is approximately normal with mean p and variance $p(1-p)/n$, if p is not too close to either 0 or 1 and if n is relatively large. Typically, to apply this approximation we require that np and $n(1-p)$ be greater than or equal to 5. We will use the normal approximation in this section.

Normal Approximation for a Binomial Proportion

If n is large, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately standard normal.

To construct the confidence interval on p, note that

$$P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) \simeq 1 - \alpha$$

so

$$P\left(-z_{\alpha/2} \leq \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) \simeq 1-\alpha$$

This may be rearranged as

$$P\left(\hat{P}-z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{P}+z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \simeq 1-\alpha \qquad (8\text{-}21)$$

The quantity $\sqrt{p(1-p)/n}$ in Equation 8-21 is called the *standard error of the point estimator* $\hat{P}$. This was discussed in Chapter 7. Unfortunately, the upper and lower limits of the confidence interval obtained from Equation 8-21 contain the unknown parameter p. However, as suggested at the end of Section 8-1.5, a solution that is often safisfactory is to replace p by $\hat{P}$ in the standard error, which results in

$$P\left(\hat{P}-z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq p \leq \hat{P}+z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) \simeq 1-\alpha \qquad (8\text{-}22)$$

This leads to the approximate $100(1-\alpha)\%$ confidence interval on p.

Approximate Confidence Interval on a Binomial Proportion

If $\hat{p}$ is the proportion of observations in a random sample of size n that belongs to a class of interest, an approximate $100(1-\alpha)\%$ confidence interval on the proportion p of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (8\text{-}23)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

This procedure depends on the adequacy of the normal approximation to the binomial. To be reasonably conservative, this requires that np and $n(1-p)$ be greater than or equal to 5. In situations when this approximation is inappropriate, particularly in cases when n is small, other methods must be used. Tables of the binomial distribution could be used to obtain a confidence interval for p. However, we could also use numerical methods that are implemented on the binomial probability mass function in some computer program.

8-5 Guidelines for Constructing Confidence Intervals

The most difficult step in constructing a confidence interval is often the match of the appropriate calculation to the objective of the study. Common cases are listed in Table 8-1 along with the reference to the section that covers the appropriate calculation for a confidence interval test. Table 8-1 provides a simple road map to help select the appropriate analysis. Two primary comments can help identify the analysis:

1. Determine the parameter (and the distribution of the data) that will be bounded by the confidence interval or tested by the hypothesis.
2. Check if other parameters are known or need to be estimated.

TABLE • 8-1 The Roadmap for Constructing Confidence Intervals and Performing Hypothesis Tests, One-Sample Case

Parameter to Be Bounded by the Confidence Interval or Tested with a Hypothesis?	Symbol	Other Parameters?	Confidence Interval Section	Hypothesis Test Section	Comments
Mean of normal distribution	μ	Standard deviation σ known	8-1	9-2	Large sample size is often taken to be $n \geq 40$
Mean of arbitrary distribution with large sample size	μ	Sample size large enough that central limit theorem applies and σ is essentially known	8-1.5	9-2.5	
Mean of normal distribution	μ	Standard deviation σ unknown and estimated	8-2	9-3	
Variance (or standard deviation) of normal distribution	σ^2	Mean μ unknown and estimated	8-3	9-4	
Population proportion	p	None	8-4	9-5	

In Chapter 9, we will study a procedure closely related to confidence intervals called *hypothesis testing*. Table 8-1 can be used for those procedures also. This road map will be extended to more cases in Chapter 10.

Content is copied from book: Applied Statistics and Probability for Engineers