6

Introduction to Formal Statistical Inference

Formal statistical inference uses probability theory to quantify the reliability of data-based conclusions. This chapter introduces the logic involved in several general types of formal statistical inference. Then the most common specific methods for one- and two-sample statistical studies are discussed.

The chapter begins with an introduction to confidence interval estimation, using the important case of large-sample inference for a mean. Then the topic of significance testing is considered, again using the case of large-sample inference for a mean. With the general notions in hand, successive sections treat the standard one- and two-sample confidence interval and significance-testing methods for means, then variances, and then proportions. Finally, the important topics of tolerance and prediction intervals are introduced.

6.1 Large-Sample Confidence Intervals for a Mean

Many important engineering applications of statistics fit the following standard mold. Values for parameters of a data-generating process are unknown. Based on data, the object is

1. identify an interval of values likely to contain an unknown parameter (or a function of one or more parameters) and
2. quantify "how likely" the interval is to cover the correct value.

For example, a piece of equipment that dispenses baby food into jars might produce an unknown mean fill level, μ. Determining a data-based interval likely to

contain μ and an evaluation of the reliability of the interval might be important. Or a machine that puts threads on U-bolts might have an inherent variation in thread lengths, describable in terms of a standard deviation, σ. The point of data collection might then be to produce an interval of likely values for σ, together with a statement of how reliable the interval is. Or two different methods of running a pelletizing machine might have different unknown propensities to produce defective pellets, (say, p_1 and p_2). A data-based interval for $p_1 - p_2$, together with an associated statement of reliability, might be needed.

The type of formal statistical inference designed to deal with such problems is called **confidence interval estimation**.

Definition 1

A **confidence interval** for a parameter (or function of one or more parameters) is a data-based interval of numbers thought likely to contain the parameter (or function of one or more parameters) possessing a stated probability-based *confidence* or reliability.

This section discusses how basic probability facts lead to simple large-sample formulas for confidence intervals for a mean, μ. The unusual case where the standard deviation σ is known is treated first. Then parallel reasoning produces a formula for the much more common situation where σ is not known. The section closes with discussions of three practical issues in the application of confidence intervals.

6.1.1 A Large-n Confidence Interval for μ Involving σ

The final example in Section 5.5 involved a physically stable filling process known to have a net weight standard deviation of $\sigma = 1.6$ g. Since, for large n, the sample mean of iid random variables is approximately normal, Example 26 of Chapter 5 argued that for $n = 47$ and

$$\bar{x} = \text{the sample mean net fill weight of 47 jars filled by the process (g)}$$

there is an approximately 80% chance that $\bar{x}$ is within .3 gram of μ. This fact is pictured again in Figure 6.1.

Notational conventions

We need to interrupt for a moment to discuss notation. In Chapter 5, capital letters were carefully used as symbols for random variables and corresponding lowercase letters for their possible or observed values. But here a lowercase symbol, $\bar{x}$, has been used for the sample mean *random variable*. This is fairly standard statistical usage, and it is in keeping with the kind of convention used in Chapters 3 and 4. We are thus going to now abandon strict adherence to the capitalization convention introduced in Chapter 5. Random variables will often be symbolized using lowercase letters and the same symbols used for their observed values. The Chapter 5 capitalization convention is especially helpful in learning the basics of probability. But once those basics are mastered, it is common to abuse notation and

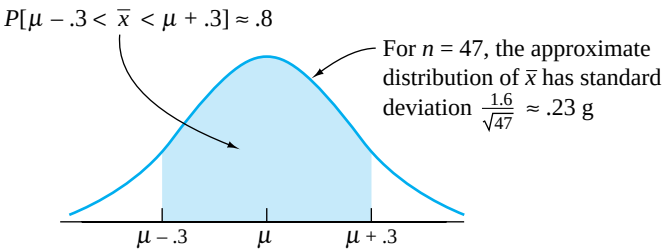

Figure 6.1 Approximate probability distribution for $\bar{x}$ based on $n = 47$

to determine from context whether a random variable or its observed value is being discussed.

The most common way of thinking about a graphic like Figure 6.1 is to think of the possibility that

$$\mu - .3 < \bar{x} < \mu + .3 \qquad \textbf{(6.1)}$$

in terms of whether or not $\bar{x}$ falls in an interval of length $2(.3) = .6$ centered at μ. But the equivalent is to consider whether or not an interval of length .6 centered at $\bar{x}$ falls on top of μ. Algebraically, inequality (6.1) is equivalent to

$$\bar{x} - .3 < \mu < \bar{x} + .3 \qquad \textbf{(6.2)}$$

which shifts attention to this second way of thinking. The fact that expression (6.2) has about an 80% chance of holding true anytime a sample of 47 fill weights is taken suggests that the *random interval*

$$(\bar{x} - .3, \bar{x} + .3) \qquad \textbf{(6.3)}$$

might be used as a confidence interval for μ, with 80% associated reliability or confidence.

Example 1

A Confidence Interval for a Process Mean Fill Weight

Suppose a sample of $n = 47$ jars produces $\bar{x} = 138.2$ g. Then expression (6.3) suggests that the interval with endpoints

$$138.2 \text{ g} \pm .3 \text{ g}$$

(i.e., the interval from 137.9 g to 138.5 g) be used as an 80% confidence interval for the process mean fill weight.

It is not hard to generalize the logic that led to expression (6.3). Anytime an iid model is appropriate for the elements of a large sample, the central limit theorem implies that the sample mean $\bar{x}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Then, if for $p > .5$, z is the p quantile of the standard normal distribution, the probability that

$$\mu - z\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + z\frac{\sigma}{\sqrt{n}} \tag{6.4}$$

is approximately $1 - 2(1 - p)$. But inequality (6.4) can be rewritten as

$$\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\frac{\sigma}{\sqrt{n}} \tag{6.5}$$

and thought of as the eventuality that the random interval with endpoints

Large-sample known σ confidence limits for μ

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \tag{6.6}$$

brackets the unknown μ. So an interval with endpoints (6.6) is an approximate confidence interval for μ (with confidence level $1 - 2(1 - p)$).

In an application, z in equation (6.6) is chosen so that the standard normal probability between $-z$ and z corresponds to a desired confidence level. Table 3.10 (of standard normal quantiles) on page 89 or Table B.3 (of standard normal cumulative probabilities) can be used to verify the appropriateness of the entries in Table 6.1. (This table gives values of z for use in expression (6.6) for some common confidence levels.)

Table 6.1
z's for Use in Two-sided Large-*n* Intervals for μ

Desired Confidence	z
80%	1.28
90%	1.645
95%	1.96
98%	2.33
99%	2.58

Example 2

Confidence Interval for the Mean Deviation from Nominal in a Grinding Operation

Dib, Smith, and Thompson studied a grinding process used in the rebuilding of automobile engines. The natural short-term variability associated with the diameters of rod journals on engine crankshafts ground using the process was on the order of $\sigma = .7 \times 10^{-4}$ in. Suppose that the rod journal grinding process can be thought of as physically stable over runs of, say, 50 journals or less. Then if 32 consecutive rod journal diameters have mean deviation from nominal of $\bar{x} = -.16 \times 10^{-4}$ in., it is possible to apply expression (6.6) to make a confidence interval for the current process mean deviation from nominal. Consider a 95% confidence level. Consulting Table 6.1 (or otherwise, realizing that 1.96 is the $p = .975$ quantile of the standard normal distribution), $z = 1.96$ is called for in formula (6.6) (since $.95 = 1 - 2(1 - .975)$). Thus, a 95% confidence interval for the current process mean deviation from nominal journal diameter has endpoints

$$-.16 \times 10^{-4} \pm (1.96)\frac{.7 \times 10^{-4}}{\sqrt{32}}$$

that is, endpoints

$$-.40 \times 10^{-4} \text{ in.} \quad \text{and} \quad .08 \times 10^{-4} \text{ in.}$$

An interval like this one could be of engineering importance in determining the advisability of making an adjustment to the process aim. The interval includes both positive and negative values. So although $\bar{x} < 0$, the information in hand doesn't provide enough precision to tell with any certainty in which direction the grinding process should be adjusted. This, coupled with the fact that potential machine adjustments are probably much coarser than the best-guess misadjustment of $\bar{x} = -.16 \times 10^{-4}$ in., speaks strongly against making a change in the process aim based on the current data.

6.1.2 A Generally Applicable Large-*n* Confidence Interval for μ

Although expression (6.6) provides a mathematically correct confidence interval, the appearance of σ in the formula severely limits its practical usefulness. It is unusual to have to estimate a mean μ when the corresponding σ is known (and can therefore be plugged into a formula). These situations occur primarily in manufacturing situations like those of Examples 1 and 2. Considerable past experience can sometimes give a sensible value for σ, while physical process drifts over time can put the current value of μ in question.

Happily, modification of the line of reasoning that led to expression (6.6) produces a confidence interval formula for μ that depends only on the characteristics of

a sample. The argument leading to formula (6.6) depends on the fact that for large n, $\bar{x}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$—i.e., that

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{6.7}$$

is approximately standard normal. The appearance of σ in expression (6.7) is what leads to its appearance in the confidence interval formula (6.6). But a slight generalization of the central limit theorem guarantees that for large n,

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{6.8}$$

is also approximately standard normal. And the variable (6.8) doesn't involve σ.

Beginning with the fact that (when an iid model for observations is appropriate and n is large) the variable (6.8) is approximately standard normal, the reasoning is much as before. For a positive z,

$$-z < \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} < z$$

is equivalent to

$$\mu - z\frac{s}{\sqrt{n}} < \bar{x} < \mu + z\frac{s}{\sqrt{n}}$$

which in turn is equivalent to

$$\bar{x} - z\frac{s}{\sqrt{n}} < \mu < \bar{x} + z\frac{s}{\sqrt{n}}$$

Thus, the interval with random center $\bar{x}$ and random length $2zs/\sqrt{n}$—i.e., with random endpoints

Large-sample confidence limits for μ

$$\boxed{\bar{x} \pm z\frac{s}{\sqrt{n}}} \tag{6.9}$$

can be used as an approximate confidence interval for μ. For a desired confidence, z should be chosen such that the standard normal probability between $-z$ and z corresponds to that confidence level.

Example 3

Breakaway Torques and Hard Disk Failures

F. Willett, in the article "The Case of the Derailed Disk Drives" (*Mechanical Engineering*, 1988), discusses a study done to isolate the cause of "blink code A failure" in a model of Winchester hard disk drive. Included in that article are the data given in Figure 6.2. These are breakaway torques (units are inch ounces) required to loosen the drive's interrupter flag on the stepper motor shaft for 26 disk drives returned to the manufacturer for blink code A failure. For these data, $\bar{x} = 11.5$ in. oz and $s = 5.1$ in. oz.

0	0 2 3
0	7 8 8 9 9
1	0 0 0 1 1 2 2 2 3
1	5 5 6 6 7 7 7 9
2	0
2	

Figure 6.2 Torques required to loosen 26 interrupter flags

If the disk drives that produced the data in Figure 6.2 are thought of as representing the population of drives subject to blink code A failure, it seems reasonable to use an iid model and formula (6.9) to estimate the population mean breakaway torque. Choosing to make a 90% confidence interval for μ, $z = 1.645$ is indicated in Table 6.1. And using formula (6.9), endpoints

$$11.5 \pm 1.645\frac{5.1}{\sqrt{26}}$$

(i.e., endpoints 9.9 in. oz and 13.1 in. oz) are indicated.

The interval shows that the mean breakaway torque for drives with blink code A failure was substantially below the factory's 33.5 in. oz target value. Recognizing this turned out to be key in finding and eliminating a design flaw in the drives.

6.1.3 Some Additional Comments Concerning Confidence Intervals

Formulas (6.6) and (6.9) have been used to make confidence statements of the type "μ is between a and b." But often a statement like "μ is at least c" or "μ is no more than d" would be of more practical value. For example, an automotive engineer might wish to state, "The mean NO emission for this engine is at most 5 ppm." Or a civil engineer might want to make a statement like "the mean compressive

strength for specimens of this type of concrete is at least 4188 psi." That is, practical engineering problems are sometimes best addressed using one-sided confidence intervals.

Making one-sided intervals

There is no real problem in coming up with formulas for one-sided confidence intervals. If you have a workable two-sided formula, all that must be done is to

1. replace the lower limit with $-\infty$ or the upper limit with $+\infty$ and
2. adjust the stated confidence level appropriately upward (this usually means dividing the "unconfidence level" by 2).

This prescription works not only with formulas (6.6) and (6.9) but also with the rest of the two-sided confidence intervals introduced in this chapter.

Example 3 (*continued*)

For the mean breakaway torque for defective disk drives, consider making a one-sided 90% confidence interval for μ of the form $(-\infty, \#)$, for # an appropriate number. Put slightly differently, consider finding a 90% *upper confidence bound* for μ, (say, #).

Beginning with a two-sided 80% confidence interval for μ, the lower limit can be replaced with $-\infty$ and a one-sided 90% confidence interval determined. That is, using formula (6.9), a 90% upper confidence bound for the mean breakaway torque is

$$\bar{x} + 1.28\frac{s}{\sqrt{n}} = 11.5 + 1.28\frac{5.1}{\sqrt{26}} = 12.8 \text{ in. oz}$$

Equivalently, a 90% one-sided confidence interval for μ is $(-\infty, 12.8)$.

The 12.8 in. oz figure here is less than (and closer to the sample mean than) the 13.1 in. oz upper limit from the 90% two-sided interval found earlier. In the one-sided case, $-\infty$ is declared as a lower limit so there is no risk of producing an interval containing only numbers larger than the unknown μ. Thus an upper limit smaller than that for a corresponding two-sided interval can be used.

Interpreting a confidence level

A second issue in the application of confidence intervals is a correct understanding of the technical meaning of the term *confidence*. Unfortunately, there are many possible misunderstandings. So it is important to carefully lay out what confidence does and doesn't mean.

Prior to selecting a sample and plugging into a formula like (6.6) or (6.9), the meaning of a confidence level is obvious. Choosing a (two-sided) 90% confidence level and thus $z = 1.645$ for use in formula (6.9), before the fact of sample selection and calculation, "there is about a 90% chance of winding up with an interval that brackets μ." In symbols, this might be expressed as

$$P\left[\bar{x} - 1.645\frac{s}{\sqrt{n}} < \mu < \bar{x} + 1.645\frac{s}{\sqrt{n}}\right] \approx .90$$

But how to think about a confidence level *after* sample selection? This is an entirely different matter. Once numbers have been plugged into a formula like (6.6) or (6.9), the die has already been cast, and the numerical interval is either right or wrong. The practical difficulty is that while which is the case can't be determined, it no longer makes logical sense to attach a probability to the correctness of the interval. For example, it would make no sense to look again at the two-sided interval found in Example 3 and try to say something like "there is a 90% probability that μ is between 9.9 in. oz and 13.1 in. oz." μ is not a random variable. It is a fixed (although unknown) quantity that either is or is not between 9.9 and 13.1. There is no probability left in the situation to be discussed.

So what does it mean that (9.9, 13.1) is a 90% confidence interval for μ? Like it or not, the phrase "90% confidence" refers more to the method used to obtain the interval (9.9, 13.1) than to the interval itself. In coming up with the interval, methodology has been used that would produce numerical intervals bracketing μ in about 90% of repeated applications. But the effectiveness of the particular interval in this application is unknown, and it is not quantifiable in terms of a probability. A person who (in the course of a lifetime) makes many 90% confidence intervals can expect to have a "lifetime success rate" of about 90%. But the effectiveness of any particular application will typically be unknown.

A short statement summarizing this discussion as "the authorized interpretation of confidence" will be useful.

Definition 2 (*Interpretation of a Confidence Interval*)

To say that a numerical interval (a, b) is (for example) a 90% confidence interval for a parameter is to say that in obtaining it, one has applied methods of data collection and calculation that would produce intervals bracketing the parameter in about 90% of repeated applications. Whether or not the particular interval (a, b) brackets the parameter is unknown and not describable in terms of a probability.

The reader may feel that the statement in Definition 2 is a rather weak meaning for the reliability figure associated with a confidence interval. Nevertheless, the statement in Definition 2 is the correct interpretation and is all that can be rationally expected. And despite the fact that the correct interpretation may initially seem somewhat unappealing, confidence interval methods have proved themselves to be of great practical use.

Sample sizes for estimating μ

As a final consideration in this introduction to confidence intervals, note that formulas like (6.6) and (6.9) can give some crude quantitative answers to the question, "How big must n be?" Using formula (6.9), for example, if you have in mind (1) a desired confidence level, (2) a worst-case expectation for the sample standard deviation, and (3) a desired precision of estimation for μ, it is a simple matter to solve for a corresponding sample size. That is, suppose that the desired confidence level dictates the use of the value z in formula (6.9), s is some likely worst-case

value for the sample standard deviation, and you want to have confidence limits (or a limit) of the form $\bar{x} \pm \Delta$. Setting

$$\Delta = z\frac{s}{\sqrt{n}}$$

and solving for n produces the requirement

$$n = \left(\frac{zs}{\Delta}\right)^2$$

Example 3 (*continued*)

Suppose that in the disk drive problem, engineers plan to follow up the analysis of the data in Figure 6.2 with the testing of a number of new drives. This will be done after subjecting them to accelerated (high) temperature conditions, in an effort to understand the mechanism behind the creation of low breakaway torques. Further suppose that the mean breakaway torque for temperature-stressed drives is to be estimated with a two-sided 95% confidence interval and that the torque variability expected in the new temperature-stressed drives is no worse than the $s = 5.1$ in. oz figure obtained from the returned drives. A ± 1 in. oz precision of estimation is desired. Then using the plus-or-minus part of formula (6.9) and remembering Table 6.1, the requirement is

$$1 = 1.96\frac{5.1}{\sqrt{n}}$$

which, when solved for n, gives

$$n = \left(\frac{(1.96)(5.1)}{1}\right)^2 \approx 100$$

A study involving in the neighborhood of $n = 100$ temperature-stressed new disk drives is indicated. If this figure is impractical, the calculations at least indicate that dropping below this sample size will (unless the variability associated with the stressed new drives is less than that of the returned drives) force a reduction in either the confidence or the precision associated with the final interval.

For two reasons, the kind of calculations in the previous example give somewhat less than an ironclad answer to the question of sample size. The first is that they are only as good as the prediction of the sample standard deviation, s. If s is underpredicted, an n that is not really large enough will result. (By the same token, if one is excessively conservative and overpredicts s, an unnecessarily large sample size will result.) The second issue is that expression (6.9) remains a large-sample formula. If calculations like the preceding ones produce n smaller than, say, 25 or 30, the value should be increased enough to guarantee that formula (6.9) can be applied.

Section 1 Exercises

1. Interpret the statement, "The interval from 6.3 to 7.9 is a 95% confidence interval for the mean μ."
2. In Chapter Exercise 2 of Chapter 3, there is a data set consisting of the aluminum contents of 26 bihourly samples of recycled PET plastic from a recycling facility. Those 26 measurements have $\bar{y} = 142.7$ ppm and $s \approx 98.2$ ppm. Use these facts to respond to the following. (Assume that $n = 26$ is large enough to permit the use of large-sample formulas in this case.)
   (a) Make a 90% two-sided confidence interval for the mean aluminum content of such specimens over the 52-hour study period.
   (b) Make a 95% two-sided confidence interval for the mean aluminum content of such specimens over the 52-hour study period. How does this compare to your answer to part (a)?
   (c) Make a 90% upper confidence bound for the mean aluminum content of such samples over the 52-hour study period. (Find # such that $(-\infty, \#)$ is a 90% confidence interval.) How does this value compare to the upper endpoint of your interval from part (a)?
   (d) Make a 95% upper confidence bound for the mean aluminum content of such samples over the 52-hour study period. How does this value compare to your answer to part (c)?
   (e) Interpret your interval from (a) for someone with little statistical background. (Speak in the context of the recycling study and use Definition 2 as your guide.)
3. Return to the context of Exercise 2. Suppose that in order to monitor for possible process changes, future samples of PET will be taken. If it is desirable to estimate the mean aluminum content with ± 20 ppm precision and 90% confidence, what future sample size do you recommend?
4. DuToit, Hansen, and Osborne measured the diameters of some no. 10 machine screws with two different calipers (digital and vernier scale). Part of their data are recorded here. Given in the small frequency table are the measurements obtained on 50 screws by one of the students using the digital calipers.

Diameter (mm)	Frequency
4.52	1
4.66	4
4.67	7
4.68	7
4.69	14
4.70	9
4.71	4
4.72	4

   (a) Compute the sample mean and standard deviation for these data.
   (b) Use your sample values from (a) and make a 98% two-sided confidence interval for the mean diameter of such screws as measured by this student with these calipers.
   (c) Repeat part (b) using 99% confidence. How does this interval compare with the one from (b)?
   (d) Use your values from (a) and find a 98% lower confidence bound for the mean diameter. (Find a number # such that $(\#, \infty)$ is a 98% confidence interval.) How does this value compare to the lower endpoint of your interval from (b)?
   (e) Repeat (d) using 99% confidence. How does the value computed here compare to your answer to (d)?
   (f) Interpret your interval from (b) for someone with little statistical background. (Speak in the context of the diameter measurement study and use Definition 2 as your guide.)

6.2 Large-Sample Significance Tests for a Mean

The last section illustrated how probability can enable confidence interval estimation. This section makes a parallel introduction of significance testing.

The goal of significance testing

Significance testing amounts to using data to quantitatively assess the plausibility of a trial value of a parameter (or function of one or more parameters). This trial value typically embodies a status quo/"pre-data" view. For example, a process engineer might employ significance testing to assess the plausibility of an ideal value of 138 g as the current process mean fill level of baby food jars. Or two different methods of running a pelletizing machine might have unknown propensities to produce defective pellets, (say, p_1 and p_2), and significance testing could be used to assess the plausibility of $p_1 - p_2 = 0$—i.e., that the two methods are equally effective.

This section describes how basic probability facts lead to simple large-sample significance tests for a mean, μ. It introduces significance testing terminology in the case where the standard deviation σ is known. Next, a five-step format for summarizing significance testing is presented. Then the more common situation of significance testing for μ where σ is not known is considered. The section closes with two discussions about practical issues in the application of significance-testing logic.

6.2.1 Large-n Significance Tests for μ Involving σ

Recall once more Example 26 in Chapter 5, where a physically stable filling process is known to have $\sigma = 1.6$ g for net weight. Suppose further that with a declared (label) weight of 135 g, process engineers have set a target mean net fill weight at $135 + 3\sigma = 139.8$ g. Finally, suppose that in a routine check of filling-process performance, intended to detect any change of the process mean from its target value, a sample of $n = 25$ jars produces $\bar{x} = 139.0$ g. What does this value have to say about the plausibility of the current process mean actually being at the target of 139.8 g?

The central limit theorem can be called on here. If indeed the current process mean is at 139.8 g, $\bar{x}$ has an approximately normal distribution with mean 139.8 g and standard deviation $\sigma/\sqrt{n} = 1.6/\sqrt{25} = .32$ g, as pictured in Figure 6.3 along with the observed value of $\bar{x} = 139.0$ g.

Figure 6.4 shows the standard normal picture that corresponds to Figure 6.3. It is based on the fact that if the current process mean is on target at 139.8 g, then the fact that $\bar{x}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n} = .32$ g implies that

$$Z = \frac{\bar{x} - 139.8}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - 139.8}{.32} \qquad \textbf{(6.10)}$$

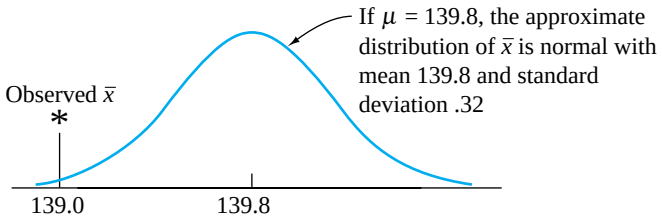

Figure 6.3 Approximate probability distribution for $\bar{x}$ if $\mu = 139.8$, and the observed value of $\bar{x} = 139.0$

is approximately standard normal. The observed $\bar{x} = 139.0$ g in Figure 6.3 has corresponding observed $z = -2.5$ in Figure 6.4.

It is obvious from either Figure 6.3 or Figure 6.4 that if the process mean is on target at 139.8 g (and thus the figures are correct), a fairly extreme/rare $\bar{x}$, or equivalently z, has been observed. Of course, extreme/rare things occasionally happen. But the nature of the observed $\bar{x}$ (or z) might instead be considered as making the possibility that the process is on target implausible.

The figures even suggest a way of quantifying their own implausibility—through calculating a probability associated with values of $\bar{x}$ (or Z) at least as extreme as the one actually observed. Now "at least as extreme" must be defined in relation to the original purpose of data collection—to detect either a decrease of μ below target or an increase above target. Not only are values $\bar{x} \le 139.0$ g ($z \le -2.5$) as extreme as that observed but so also are values $\bar{x} \ge 140.6$ g ($z \ge 2.5$). (The first kind of $\bar{x}$ suggests a decrease in μ, and the second suggests an increase.) That is, the implausibility of being on target might be quantified by noting that if this were so, only a fraction

$$\Phi(-2.5) + \big(1 - \Phi(2.5)\big) = .01$$

of all samples would produce a value of $\bar{x}$ (or Z) as extreme as the one actually observed. Put in those terms, the data seem to speak rather convincingly against the process being on target.

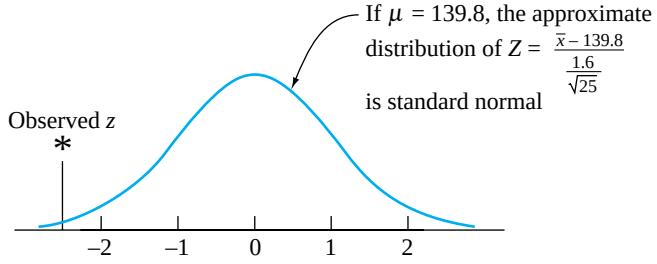

Figure 6.4 The standard normal picture corresponding to Figure 6.3

The argument that has just been made is an application of typical significance-testing logic. In order to make the pattern of thought obvious, it is useful to isolate some elements of it in definition form. This is done next, beginning with a formal restatement of the overall purpose.

Definition 3

Statistical **significance testing** is the use of data in the quantitative assessment of the plausibility of some trial value for a parameter (or function of one or more parameters).

Logically, significance testing begins with the specification of the trial or hypothesized value. Special jargon and notation exist for the statement of this value.

Definition 4

A **null hypothesis** is a statement of the form

$$\text{Parameter} = \#$$

or

$$\text{Function of parameters} = \#$$

(for some number, #) that forms the basis of investigation in a significance test. A null hypothesis is usually formed to embody a status quo/"pre-data" view of the parameter (or function of the parameter(s)). It is typically denoted as H_0.

The notion of a null hypothesis is so central to significance testing that it is common to use the term **hypothesis testing** in place of *significance testing*. The "null" part of the phrase "null hypothesis" refers to the fact that null hypotheses are statements of *no difference*, or equality. For example, in the context of the filling operation, standard usage would be to write

$$H_0: \mu = 139.8 \tag{6.11}$$

meaning that there is no difference between μ and the target value of 139.8 g.

After formulating a null hypothesis, what kinds of departures from it are of interest must be specified.

Definition 5

An **alternative hypothesis** is a statement that stands in opposition to the null hypothesis. It specifies what forms of departure from the null hypothesis are of concern. An alternative hypothesis is typically denoted as H_a. It is of the

same form as the corresponding null hypothesis, except that the equality sign is replaced by $\neq$, $>$, or $<$.

Often, the alternative hypothesis is based on an investigator's suspicions and/or hopes about the true state of affairs, amounting to a kind of *research hypothesis* that the investigator hopes to establish. For example, if an engineer tests what is intended to be a device for improving automotive gas mileage, a null hypothesis expressing "no mileage change" and an alternative hypothesis expressing "mileage improvement" would be appropriate.

Definitions 4 and 5 together imply that for the case of testing about a single mean, the three possible pairs of null and alternative hypotheses are

$$\begin{array}{lll} H_0: \mu = \# & H_0: \mu = \# & H_0: \mu = \# \\ H_a: \mu > \# & H_a: \mu < \# & H_a: \mu \neq \# \end{array}$$

In the example of the filling operation, there is a need to detect both the possibility of consistently underfilled ($\mu < 139.8$ g) and the possibility of consistently overfilled ($\mu > 139.8$ g) jars. Thus, an appropriate alternative hypothesis is

$$H_a: \mu \neq 139.8 \qquad \textbf{(6.12)}$$

Once null and alternative hypotheses have been established, it is necessary to lay out carefully how the data will be used to evaluate the plausibility of the null hypothesis. This involves specifying a statistic to be calculated, a probability distribution appropriate for it if the null hypothesis is true, and what kinds of observed values will make the null hypothesis seem implausible.

Definition 6

A **test statistic** is the particular form of numerical data summarization used in a significance test. The formula for the test statistic typically involves the number appearing in the null hypothesis.

Definition 7

A **reference (or null) distribution** for a test statistic is the probability distribution describing the test statistic, provided the null hypothesis is in fact true.

The values of the test statistic considered to cast doubt on the validity of the null hypothesis are specified after looking at the form of the alternative hypothesis. Roughly speaking, values are identified that are more likely to occur if the alternative hypothesis is true than if the null hypothesis holds.

The discussion of the filling process scenario has vacillated between using $\bar{x}$ and its standardized version Z given in equation (6.10) for a test statistic. Equation (6.10) is a specialized form of the general (large-n, known σ) test statistic for μ,

Large-sample known σ test statistic for μ

$$Z = \frac{\bar{x} - \#}{\frac{\sigma}{\sqrt{n}}} \tag{6.13}$$

for the present scenario, where the hypothesized value of μ is 139.8, $n = 25$, and $\sigma = 1.6$. It is most convenient to think of the test statistic for this kind of problem in the standardized form shown in equation (6.13) rather than as $\bar{x}$ itself. Using form (6.13), the reference distribution will always be the same—namely, standard normal.

Continuing with the filling example, note that if instead of the null hypothesis (6.11), the alternative hypothesis (6.12) is operating, observed $\bar{x}$'s much larger or much smaller than 139.8 will tend to result. Such $\bar{x}$'s will then, via equation (6.13), translate respectively to large or small (that is, large negative numbers in this case) observed values of Z—i.e., large values $|z|$. Such observed values render the null hypothesis implausible.

Having specified how data will be used to judge the plausibility of the null hypothesis, it remains to collect them, plug them into the formula for the test statistic, and (using the calculated value and the reference distribution) arrive at a quantitative assessment of the plausibility of H_0. There is jargon for the form this will take.

Definition 8

The **observed level of significance** or ***p*-value** in a significance test is the probability that the reference distribution assigns to the set of possible values of the test statistic that are at least as extreme as the one actually observed (in terms of casting doubt on the null hypothesis).

Small p-values are evidence against H_0

The smaller the observed level of significance, the stronger the evidence against the validity of the null hypothesis. In the context of the filling operation, with an observed value of the test statistic of

$$z = -2.5$$

the p-value or observed level of significance is

$$\Phi(-2.5) + \big(1 - \Phi(2.5)\big) = .01$$

which gives fairly strong evidence against the possibility that the process mean is on target.

6.2.2 A Five-Step Format for Summarizing Significance Tests

Five-step significance testing format

It is helpful to lay down a step-by-step format for organizing write-ups of significance tests. The one that will be used in this text includes the following five steps:

Step 1 State the null hypothesis.

Step 2 State the alternative hypothesis.

Step 3 State the test criteria. That is, give the formula for the test statistic (plugging in only a hypothesized value from the null hypothesis, but not any sample information) and the reference distribution. Then state in general terms what observed values of the test statistic will constitute evidence against the null hypothesis.

Step 4 Show the sample-based calculations.

Step 5 Report an observed level of significance and (to the extent possible) state its implications in the context of the real engineering problem.

Example 4

A Significance Test Regarding a Process Mean Fill Level

The five-step significance-testing format can be used to write up the preceding discussion of the filling process.

1. $H_0: \mu = 139.8$.
2. $H_a: \mu \neq 139.8$.
3. The test statistic is
$$Z = \frac{\bar{x} - 139.8}{\frac{\sigma}{\sqrt{n}}}$$
The reference distribution is standard normal, and large observed values $|z|$ will constitute evidence against H_0.
4. The sample gives
$$z = \frac{139.0 - 139.8}{\frac{1.6}{\sqrt{100}}} = -2.5$$
5. The observed level of significance is
$$\begin{aligned} &P[\text{a standard normal variable} \leq -2.5] \\ &\qquad + P[\text{a standard normal variable} \geq 2.5] \\ &= P\,[|\text{a standard normal variable}| \geq 2.5] \\ &= .01 \end{aligned}$$

This is reasonably strong evidence that the process mean fill level is not on target.

6.2.3 Generally Applicable Large-n Significance Tests for μ

The significance-testing method used to carry the discussion thus far is easy to discuss and understand but of limited practical use. The problem with it is that statistic (6.13) involves the parameter σ. As remarked in Section 6.1, there are few engineering contexts where one needs to make inferences regarding μ but knows the corresponding σ. Happily, because of the same probability fact that made it possible to produce a large-sample confidence interval formula for μ free of σ, it is also possible to do large-n significance testing for μ without having to supply σ.

For observations that are describable as essentially equivalent to random selections with replacement from a single population with mean μ and variance σ^2, if n is large,

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is approximately standard normal. This means that for large n, to test

$$H_0: \mu = \#$$

a widely applicable method will simply be to use the logic already introduced but with the statistic

Large-sample test statistic for μ

$$Z = \frac{\bar{x} - \#}{\frac{s}{\sqrt{n}}} \qquad \textbf{(6.14)}$$

in place of statistic (6.13).

Example 5 (*Example 3 revisited*)

Significance Testing and Hard Disk Failures

Consider again the problem of disk drive blink code A failure. Breakaway torques set at the factory on the interrupter flag connection to the stepper motor shaft averaged 33.5 in. oz, and there was suspicion that blink code A failure was associated with reduced breakaway torque. Recall that a sample of $n = 26$ failed drives had breakaway torques (given in Figure 6.2) with $\bar{x} = 11.5$ in. oz and $s = 5.1$ in. oz.

Consider the situation of an engineer wishing to judge the extent to which the data in hand debunk the possibility that drives experiencing blink code A failure

Example 5 (*continued*)

have mean breakaway torque equal to the factory-set mean value of 33.5 in. oz. The five-step significance-testing format can be used.

1. H_0: $\mu = 33.5$.
2. H_a: $\mu < 33.5$.
   (Here the alternative hypothesis is directional, amounting to a research hypothesis based on the engineer's suspicions about the relationship between drive failure and breakaway torque.)
3. The test statistic is

$$Z = \frac{\bar{x} - 33.5}{\frac{s}{\sqrt{n}}}$$

   The reference distribution is standard normal, and small observed values z will constitute evidence against the validity of H_0. (Means less than 33.5 will tend to produce $\bar{x}$'s of the same nature and therefore small—i.e., large negative—z's.)
4. The sample gives

$$z = \frac{11.5 - 33.5}{\frac{5.1}{\sqrt{26}}} = -22.0$$

5. The observed level of significance is

$$P[\text{a standard normal variable} < -22.0] \approx 0$$

   The sample provides overwhelming evidence that failed drives have a mean breakaway torque below the factory-set level.

It is important not to make too much of a logical jump here to an incorrect conclusion that this work constitutes the complete solution to the real engineering problem. Drives returned for blink code A failure have substandard breakaway torques. But in the absence of evidence to the contrary, it is possible that they are no different in that respect from nonfailing drives currently in the field. And even if reduced breakaway torque is at fault, a real-world fix of the drive failure problem requires the identification and prevention of the physical mechanism producing it. This is not to say the significance test lacks importance, but rather to remind the reader that it is but one of many tools an engineer uses to do a job.

6.2.4 Significance Testing and Formal Statistical Decision Making (*Optional*)

The basic logic introduced in this section is sometimes applied in a decision-making context, where data are being counted on to provide guidance in choosing between two rival courses of action. In such cases, a decision-making framework is often built into the formal statistical analysis in an explicit way, and some additional terminology and patterns of thought are standard.

In some decision-making contexts, it is possible to conceive of two different possible decisions or courses of action as being related to a null and an alternative hypothesis. For example, in the filling-process scenario, H_0: $\mu = 139.8$ might correspond to the course of action "leave the process alone," and H_a: $\mu \neq 139.8$ could correspond to the course of action "adjust the process." When such a correspondence holds, two different errors are possible in the decision-making process.

Definition 9

When significance testing is used in a decision-making context, deciding in favor of H_a when in fact H_0 is true is called a **type I error**.

Definition 10

When significance testing is used in a decision-making context, deciding in favor of H_0 when in fact H_a is true is called a **type II error**.

The content of these two definitions is represented in the 2×2 table pictured in Figure 6.5. In the filling-process problem, a type I error would be adjusting an on-target process. A type II error would be failing to adjust an off-target process.

Significance testing is harnessed and used to come to a decision by choosing a critical value and, if the observed level of significance is smaller than the critical value (thus making the null hypothesis correspondingly implausible), deciding in favor of H_a. Otherwise, the course of action corresponding to H_0 is followed. The critical value for the observed level of significance ends up being the a priori

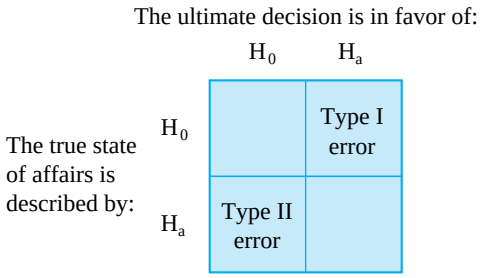

Figure 6.5 Four potential outcomes in a decision problem

probability the decision maker runs of deciding in favor of H_a, calculated supposing H_0 to be true. There is special terminology for this concept.

Definition 11

When significance testing is used in a decision-making context, a critical value separating those large observed levels of significance for which H_0 will be accepted from those small observed levels of significance for which H_0 will be rejected in favor of H_a is called the **type I error probability** or the **significance level**. The symbol α is usually used to stand for the type I error probability.

It is standard practice to use small numbers, like .1, .05, or even .01, for α. This puts some inertia in favor of H_0 into the decision-making process. (Such a practice guarantees that *type I* errors won't be made very often. But at the same time, it creates an asymmetry in the treatment of H_0 and H_a that is not always justified.)

Definition 10 and Figure 6.5 make it clear that type I errors are not the only undesirable possibility. The possibility of type II errors must also be considered.

Definition 12

When significance testing is used in a decision-making context, the probability—calculated supposing a particular parameter value described by H_a holds—that the observed level of significance is bigger than α (i.e., H_0 is not rejected) is called a **type II error probability**. The symbol β is usually used to stand for a type II error probability.

For most of the testing methods studied in this book, calculation of β's is more than the limited introduction to probability given in Chapter 5 will support. But the job can be handled for the simple known-σ situation that was used to introduce the topic of significance testing. And making a few such calculations will provide some intuition consistent with what, qualitatively at least, holds in general.

Example 4 (*continued*)

Again consider the filling process and testing $H_0: \mu = 139.8$ vs. $H_a: \mu \neq 139.8$. This time suppose that significance testing based on $n = 25$ will be used tomorrow to decide whether or not to adjust the process. Type II error probabilities, calculated supposing $\mu = 139.5$ and $\mu = 139.2$ for tests using $\alpha = .05$ and $\alpha = .2$, will be compared.

First consider $\alpha = .05$. The decision will be made in favor of H_0 if the p-value exceeds .05. That is, the decision will be in favor of the null hypothesis if the observed value of Z given in equation (6.10) (generalized in formula (6.13)) is such that

$$|z| < 1.96$$

i.e., if

$$139.8 - 1.96(.32) < \bar{x} < 139.8 + 1.96(.32)$$

i.e., if

$$139.2 < \bar{x} < 140.4 \qquad \textbf{(6.15)}$$

Now if μ described by H_a given in display (6.12) is the true process mean, $\bar{x}$ is not approximately normal with mean 139.8 and standard deviation .32, but rather approximately normal with mean μ and standard deviation .32. So for such a μ, expression (6.15) and Definition 12 show that the corresponding β will be the probability the corresponding normal distribution assigns to the possibility that $139.2 < \bar{x} < 140.4$. This is pictured in Figure 6.6 for the two means $\mu = 139.5$ and $\mu = 139.2$.

It is an easy matter to calculate z-values corresponding to $\bar{x} = 139.2$ and $\bar{x} = 140.4$ using means of 139.5 and 139.2 and a standard deviation of .32 and to consult a standard normal table in order to verify the correctness of the two β's marked in Figure 6.6.

Parallel reasoning for the situation with $\alpha = .2$ is as follows. The decision will be in favor of H_0 if the p-value exceeds .2. That is, the decision will be in favor of H_0 if $|z| < 1.28$—i.e., if

$$139.4 < \bar{x} < 140.2$$

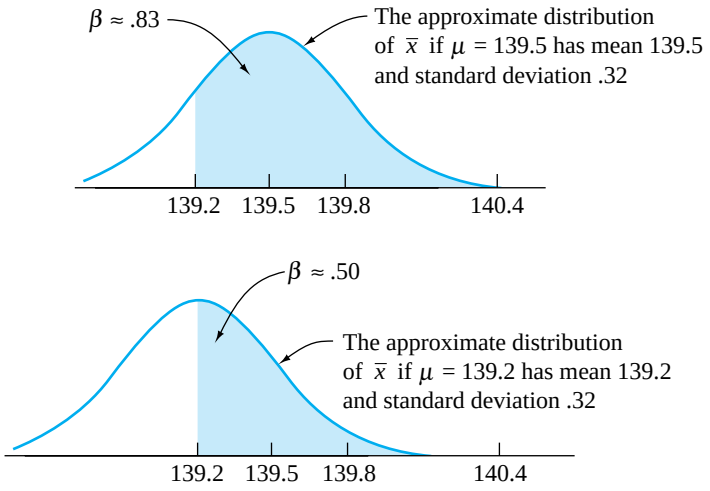

Figure 6.6 Approximate probability distributions for $\bar{x}$ for two different values of μ described by H_a and the corresponding β's, when $\alpha = .05$

Example 4 (*continued*)

If μ described by H_a is the true process mean, $\bar{x}$ is approximately normal with mean μ and standard deviation .32. So the corresponding β will be the probability this normal distribution assigns to the possibility that $139.4 < \bar{x} < 140.2$. This is pictured in Figure 6.7 for the two means $\mu = 139.5$ and $\mu = 139.2$, having corresponding type II error probabilities $\beta = .61$ and $\beta = .27$.

The calculations represented by the two figures are collected in Table 6.2. Notice two features of the table. First, the β values for $\alpha = .05$ are larger than those for $\alpha = .2$. If one wants to run only a 5% chance of (incorrectly) deciding to adjust an on-target process, the price to be paid is a larger probability of failure to recognize an off-target condition. Secondly, the β values for $\mu = 139.2$ are smaller than the β values for $\mu = 139.5$. The further the filling process is from being on target, the less likely it is that the off-target condition will fail to be detected.

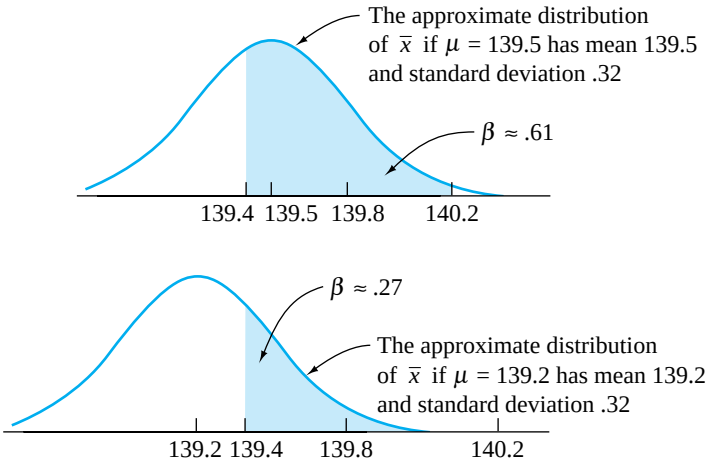

Figure 6.7 Approximate probability distributions for $\bar{x}$ for two different values of μ described by H_a and the corresponding β's, when $\alpha = .2$

Table 6.2
$n = 25$ type II error probabilities (β)

		μ	
		139.2	139.5
α	.05	.50	.83
	.2	.27	.61

The story told by Table 6.2 applies in qualitative terms to all uses of significance testing in decision-making contexts. The further H_0 is from being true, the smaller the corresponding β. And small α's imply large β's and vice versa.

The effect of sample size on β's

There is one other element of this general picture that plays an important role in the determination of error probabilities. That is the matter of sample size. If a sample size can be increased, for a given α, the corresponding β's can be reduced. Redo the calculations of the previous example, this time supposing that $n = 100$ rather than 25. Table 6.3 shows the type II error probabilities that should result, and comparison with Table 6.2 serves to indicate the sample-size effect in the filling-process example.

Analogy between testing and a criminal trial

An analogy helpful in understanding the standard logic applied when significance testing is employed in decision-making involves thinking of the process of coming to a decision as a sort of legal proceeding, like a criminal trial. In a criminal trial, there are two opposing hypotheses, namely

H_0: The defendant is innocent

H_a: The defendant is guilty

Evidence, playing a role similar to the data used in testing, is gathered and used to decide between the two hypotheses. Two types of potential error exist in a criminal trial: the possibility of convicting an innocent person (parallel to the type I error) and the possibility of acquitting a guilty person (similar to the type II error). A criminal trial is a situation where the two types of error are definitely thought of as having differing consequences, and the two hypotheses are treated asymmetrically. The a priori presumption in a criminal trial is in favor of H_0, the defendant's innocence. In order to keep the chance of a false conviction small (i.e., keep α small), overwhelming evidence is required for conviction, in much the same way that if small α is used in testing, extreme values of the test statistic are needed in order to indicate rejection of H_0. One consequence of this method of operation in criminal trials is that there is a substantial chance that a guilty individual will be acquitted, in the same way that small α's produce big β's in testing contexts.

This significance testing/criminal trial parallel is useful, but do not make more of it than is justified. Not all significance-testing applications are properly thought of in this light. And few engineering scenarios are simple enough to reduce to a "decide between H_0 and H_a" choice. Sensible applications of significance testing are

Table 6.3
$n = 100$ Type II Error Probabilities (β)

		μ	
		139.2	139.5
α	.05	.04	.53
	.2	.01	.28

often only steps of "evidence evaluation" in a many-faceted, data-based detective job necessary to solve an engineering problem. And even when a real problem can be reduced to a simple "decide between H_0 and H_a" framework, it need not be the case that the "choose a small α" logic is appropriate. In some engineering contexts, the practical consequences of a type II error are such that rational decision-making strikes a balance between the opposing goals of small α and small β's.

6.2.5 Some Comments Concerning Significance Testing and Estimation

Confidence interval estimation and significance testing are the two most commonly used forms of formal statistical inference. These having been introduced, it is appropriate to offer some comparative comments about their practical usefulness and, in the process, admit to an *estimation orientation* that will be reflected in much of the rest of this book's treatment of formal inference.

More often than not, engineers need to know "What is the value of the parameter?" rather than "Is the parameter equal to some hypothesized value?" And it is confidence interval estimation, not significance testing, that is designed to answer the first question. A confidence interval for a mean breakaway torque of from 9.9 in. oz to 13.1 in. oz says what values of μ seem plausible. A tiny observed level of significance in testing H_0: $\mu = 33.5$ says only that the data speak clearly against the possibility that $\mu = 33.5$, but it doesn't give any clue to the likely value of μ.

"Statistical significance" and practical importance

The fact that significance testing doesn't produce any useful indication of what parameter values are plausible is sometimes obscured by careless interpretation of semistandard jargon. For example, it is common in some fields to term p-values less than .05 "statistically significant" and ones less than .01 "highly significant." The danger in this kind of usage is that "significant" can be incorrectly heard to mean "of great practical consequence" and the p-value incorrectly interpreted as a measure of how much a parameter differs from a value stated in a null hypothesis. One reason this interpretation doesn't follow is that the observed level of significance in a test depends not only on how far H_0 appears to be from being correct but on the sample size as well. Given a large enough sample size, any departure from H_0, whether of practical importance or not, can be shown to be "highly significant."

Example 6

Statistical Significance and Practical Importance in a Regulatory Agency Test

A good example of the previous points involves the newspaper article in Figure 6.8. Apparently the Pass Master manufacturer did enough physical mileage testing (used a large enough n) to produce a p-value less than .05 for testing a null hypothesis of no mileage improvement. That is, a "statistically significant" result was obtained.

But the size of the actual mileage improvement reported is only "small but real," amounting to about .8 mpg. Whether or not this improvement is of *practical importance* is a matter largely separate from the significance-testing

WASHINGTON (AP)—A gadget that cuts off a car's air conditioner when the vehicle accelerates has become the first product aimed at cutting gasoline consumption to win government endorsement.

The device, marketed under the name "Pass Master," can provide a "small but real fuel economy benefit," the Environmental Protection Agency said Wednesday.

Motorists could realize up to 4 percent fuel reduction while using their air conditioners on cars equipped with the device, the agency said. That would translate into .8-miles-per-gallon improvement for a car that normally gets 20 miles to the gallon with the air conditioner on.

The agency cautioned that the 4 percent figure was a maximum amount and could be less depending on a motorist's driving habits, the type of car and the type of air conditioner.

But still the Pass Master, which sells for less than $15, is the first of 40 products to pass the EPA's tests as making any "statistically significant" improvement in a car's mileage.

Figure 6.8 Article from *The Lafayette Journal and Courier*, Page D-3, August 28, 1980. Reprinted by permission of the Associated Press. © 1980 the Associated Press.

result. And an engineer equipped with a confidence interval for the mean mileage improvement is in a better position to judge this than is one who knows only that the p-value was less than .05.

Example 5 (*continued*)

To illustrate the effect that sample size has on observed level of significance, return to the breakaway torque problem and consider two hypothetical samples, one based on $n = 25$ and the other on $n = 100$ but both giving $\bar{x} = 32.5$ in. oz and $s = 5.1$ in. oz.

For testing H_0: $\mu = 33.5$ with H_a: $\mu < 33.5$, the first hypothetical sample gives

$$z = \frac{32.5 - 33.5}{\frac{5.1}{\sqrt{25}}} = -.98$$

with associated observed level of significance

$$\Phi(-.98) = .16$$

The second hypothetical sample gives

$$z = \frac{32.5 - 33.5}{\frac{5.1}{\sqrt{100}}} = -1.96$$

Example 5 (*continued*)

with corresponding p-value

$$\Phi(-1.96) = .02$$

Because the second sample size is larger, the second sample gives stronger evidence that the mean breakaway torque is below 33.5 in. oz. But the best data-based guess at the difference between μ and 33.5 is $\bar{x} - 33.5 = -1.0$ in. oz *in both cases*. And it is the size of the difference between μ and 33.5 that is of primary engineering importance.

It is further useful to realize that in addition to doing its primary job of providing an interval of plausible values for a parameter, a confidence interval itself also provides some significance-testing information. For example, a 95% confidence interval for a parameter contains all those values of the parameter for which significance tests using the data in hand would produce p-values bigger than 5%. (Those values not covered by the interval would have associated p-values smaller than 5%.)

Example 5 (*continued*)

Recall from Section 6.1 that a 90% one-sided confidence interval for the mean breakaway torque for failed drives is $(-\infty, 12.8)$. This means that for any value, #, larger than 12.8 in. oz, a significance test of H_0: $\mu = \#$ with H_a: $\mu < \#$ would produce a p-value less than .1. So clearly, the observed level of significance corresponding to the null hypothesis H_0: $\mu = 33.5$ is less than .1 . (In fact, as was seen earlier in this section, the p-value is 0 to two decimal places.) Put more loosely, the interval $(-\infty, 12.8)$ is a long way from containing 33.5 in. oz and therefore makes such a value of μ quite implausible.

The discussion here could well raise the question "What practical role remains for significance testing?" Some legitimate answers to this question are

1. In an almost negative way, p-values can help an engineer gauge the extent to which data in hand are inconclusive. When observed levels of significance are large, more information is needed in order to arrive at any definitive judgment.
2. Sometimes legal requirements force the use of significance testing in a compliance or effectiveness demonstration. (This was the case in Figure 6.8, where before the Pass Master could be marketed, some mileage improvement had to be legally demonstrated.)
3. There are cases where the use of significance testing in a decision-making framework is necessary and appropriate. (An example is acceptance sampling: Based on information from a sample of items from a large lot, one must determine whether or not to receive shipment of the lot.)

So, properly understood and handled, significance testing does have its place in engineering practice. Thus, although the rest of this book features estimation over significance testing, methods of significance testing will not be completely ignored.

Section 2 Exercises

1. In the aluminum contamination study discussed in Exercise 2 of Section 6.1 and in Chapter Exercise 2 of Chapter 3, it was desirable to have mean aluminum content for samples of recycled plastic below 200 ppm. Use the five-step significance-testing format and determine the strength of the evidence in the data that in fact this contamination goal has been violated. (You will want to begin with H_0: $\mu = 200$ ppm and use H_a: $\mu > 200$ ppm.)
2. Heyde, Kuebrick, and Swanson measured the heights of 405 steel punches of a particular type. These were all from a single manufacturer and were supposed to have heights of .500 in. (The stamping machine in which these are used is designed to use .500 in. punches.) The students' measurements had $\bar{x} = .5002$ in. and $s = .0026$ in. (The raw data are given in Chapter Exercise 9 of Chapter 3.)
   (a) Use the five-step format and test the hypothesis that the mean height of such punches is "on spec" (i.e., is .500 in.).
   (b) Make a 98% two-sided confidence interval for the mean height of such punches produced by this manufacturer under conditions similar to those existing when the students' punches were manufactured. Is your interval consistent with the outcome of the test in part (a)? Explain.
   (c) In the students' application, the mean height of the punches did not tell the whole story about how they worked in the stamping machine. Several of these punches had to be placed side by side and used to stamp the same piece of material. In this context, what other feature of the height distribution is almost certainly of practical importance?
3. Discuss, in the context of Exercise 2, part (a), the potential difference between statistical significance and practical importance.
4. In the context of the machine screw diameter study of Exercise 4 of Section 6.1, suppose that the nominal diameter of such screws is 4.70 mm. Use the five-step significance-testing format and assess the strength of the evidence provided by the data that the long-run mean measured diameter differs from nominal. (You will want to begin with H_0: $\mu = 4.70$ mm and use H_a: $\mu \neq 4.70$ mm.)
5. Discuss, in the context of Exercise 4, the potential difference between statistical significance and practical importance.

6.3 One- and Two-Sample Inference for Means

Sections 6.1 and 6.2 introduced the basic concepts of confidence interval estimation and significance testing. There are thousands of specific methods of these two types. This book can only discuss a small fraction that are particularly well known and useful to engineers. The next three sections consider the most elementary of these—some of those that are applicable to one- and two-sample studies—beginning in this section with methods of formal inference for means.

Inferences for a single mean, based not on the large samples of Sections 6.1 and 6.2 but instead on small samples, are considered first. In the process, it is necessary

to introduce the so-called (Student) t probability distributions. Presented next are methods of formal inference for paired data. The section concludes with discussions of both large- and small-n methods for data-based comparison of two means based on independent samples.

6.3.1 Small-Sample Inference for a Single Mean

The most important practical limitation on the use of the methods of the previous two sections is the requirement that n must be large. That restriction comes from the fact that without it, there is no way to conclude that

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{6.16}$$

is approximately standard normal. So if, for example, one mechanically uses the large-n confidence interval formula

$$\bar{x} \pm z\frac{s}{\sqrt{n}} \tag{6.17}$$

with a small sample, there is no way of assessing what actual level of confidence should be declared. That is, for small n, using $z = 1.96$ in formula (6.17) generally doesn't produce 95% confidence intervals. And without a further condition, there is neither any way to tell what confidence might be associated with $z = 1.96$ nor any way to tell how to choose z in order to produce a 95% confidence level.

There is one important special circumstance in which it is possible to reason in a way parallel to the work in Sections 6.1 and 6.2 and arrive at inference methods for means based on small sample sizes. That is the situation where it is sensible to model the observations as iid normal random variables. The normal observations case is convenient because although the variable (6.16) is not standard normal, it does have a recognized, tabled distribution. This is the **Student t distribution**.

Definition 13

The **(Student) t distribution with degrees of freedom parameter** ν is a continuous probability distribution with probability density

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \quad \text{for all } t \tag{6.18}$$

If a random variable has the probability density given by formula (6.18), it is said to have a t_ν distribution.

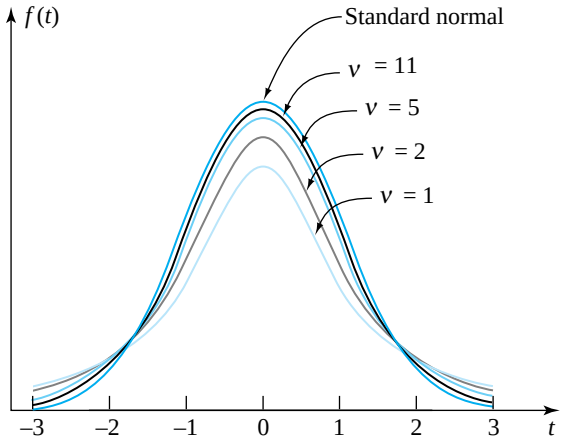

Figure 6.9 t Probability densities for $\nu = 1$, 2, 5, and 11 and the standard normal density

The word *Student* in Definition 13 was the pen name of the statistician who first came upon formula (6.18). Expression (6.18) is rather formidable looking. No direct computations with it will actually be required in this book. But, it is useful to have expression (6.18) available in order to sketch several t probability densities, to get a feel for their shape. Figure 6.9 pictures the t densities for degrees of freedom $\nu = 1$, 2, 5, and 11, along with the standard normal density.

t distributions and the standard normal distribution

The message carried by Figure 6.9 is that the t probability densities are bell shaped and symmetric about 0. They are flatter than the standard normal density but are increasingly like it as ν gets larger. In fact, for most practical purposes, for ν larger than about 30, the t distribution with ν degrees of freedom and the standard normal distribution are indistinguishable.

Probabilities for the t distributions are not typically found using the density in expression (6.18), as no simple antiderivative for $f(t)$ exists. Instead, it is common to use tables (or statistical software) to evaluate common t distribution quantiles and to get at least crude bounds on the types of probabilities needed in significance testing. Table B.4 is a typical table of t quantiles. Across the top of the table are several cumulative probabilities. Down the left side are values of the degrees of freedom parameter, ν. In the body of the table are corresponding quantiles. Notice also that the last line of the table is a "$\nu = \infty$" (i.e., standard normal) line.

Example 7

Use of a Table of t Distribution Quantiles

Suppose that T is a random variable having a t distribution with $\nu = 5$ degrees of freedom. Consider first finding the .95 quantile of T's distribution, then seeing what Table B.4 reveals about $P[T < -1.9]$ and then about $P[|T| > 2.3]$.

Example 7 (*continued*)

First, looking at the $\nu = 5$ row of Table B.4 under the cumulative probability .95, 2.015 is found in the body of the table. That is, $Q(.95) = 2.015$ or (equivalently) $P[T \leq 2.015] = .95$.

Then note that by symmetry,

$$P[T < -1.9] = P[T > 1.9] = 1 - P[T \leq 1.9]$$

Looking at the $\nu = 5$ row of Table B.4, 1.9 is between the .90 and .95 quantiles of the t_5 distribution. That is,

$$.90 < P[T \leq 1.9] \leq .95$$

so finally

$$.05 < P[T < -1.9] < .10$$

Lastly, again by symmetry,

$$\begin{aligned} P[|T| > 2.3] &= P[T < -2.3] + P[T > 2.3] = 2P[T > 2.3] \\ &= 2(1 - P[T \leq 2.3]) \end{aligned}$$

Then, from the $\nu = 5$ row of Table B.4, 2.3 is seen to be between the .95 and .975 quantiles of the t_5 distribution. That is,

$$.95 < P[T \leq 2.3] < .975$$

so

$$.05 < P[|T| > 2.3] < .10$$

The three calculations of this example are pictured in Figure 6.10.

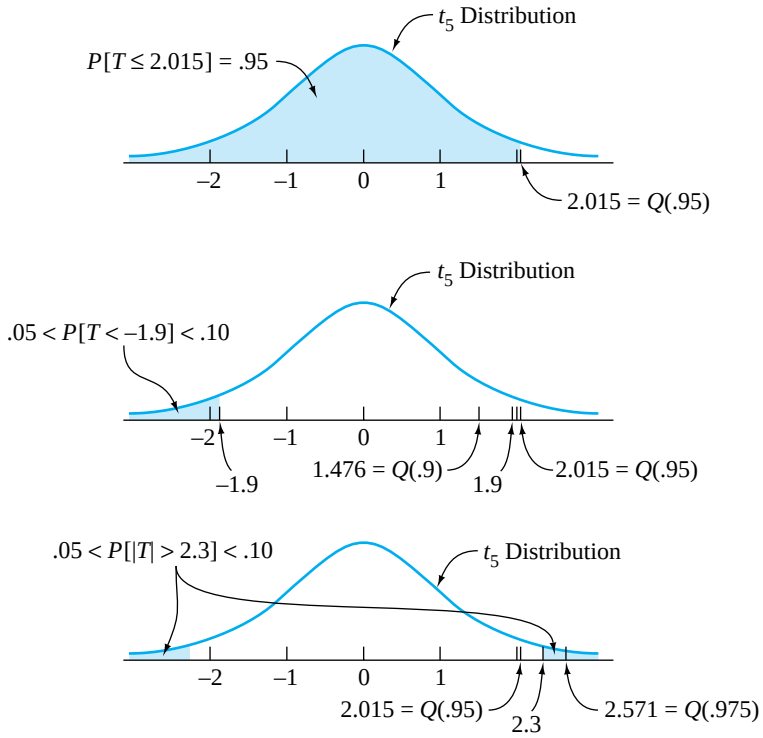

Figure 6.10 Three t_5 probability calculations for Example 7

The connection between expressions (6.18) and (6.16) that allows the development of small-n inference methods for normal observations is that if an iid normal model is appropriate,

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{6.19}$$

has the t distribution with $\nu = n - 1$ degrees of freedom. (This is consistent with the basic fact used in the previous two sections. That is, for large n, ν is large, so the t_ν distribution is approximately standard normal; and for large n, the variable (6.19) has already been treated as approximately standard normal.)

Since the variable (6.19) can under appropriate circumstances be treated as a t_{n-1} random variable, we are in a position to work in exact analogy to what was done in Sections 6.1 and 6.2 to find methods for confidence interval estimation and significance testing. That is, if a data-generating mechanism can be thought of as

essentially equivalent to drawing independent observations from a single normal distribution, a two-sided confidence interval for μ has endpoints

Normal distribution confidence limits for μ

$$\bar{x} \pm t\frac{s}{\sqrt{n}} \tag{6.20}$$

where t is chosen such that the t_{n-1} distribution assigns probability corresponding to the desired confidence level to the interval between $-t$ and t. Further, the null hypothesis

$$\mathrm{H}_0\colon \mu = \#$$

can be tested using the statistic

Normal distribution test statistic for μ

$$T = \frac{\bar{x} - \#}{\frac{s}{\sqrt{n}}} \tag{6.21}$$

and a t_{n-1} reference distribution.

Operationally, the only difference between the inference methods indicated here and the large-sample methods of the previous two sections is the exchange of standard normal quantiles and probabilities for ones corresponding to the t_{n-1} distribution. *Conceptually*, however, the nominal confidence and significance properties here are practically relevant only under the extra condition of a reasonably normal underlying distribution. Before applying either expression (6.20) or (6.21) in practice, it is advisable to investigate the appropriateness of a normal model assumption.

Example 8

Small-Sample Confidence Limits for a Mean Spring Lifetime

Part of a data set of W. Armstrong (appearing in *Analysis of Survival Data* by Cox and Oakes) gives numbers of cycles to failure of ten springs of a particular type under a stress of 950 N/mm^2. These spring-life observations are given in Table 6.4, in units of 1,000 cycles.

Table 6.4
Cycles to Failure of Ten Springs under 950 N/mm^2 Stress (10^3 cycles)

Spring Lifetimes
225, 171, 198, 189, 189
135, 162, 135, 117, 162

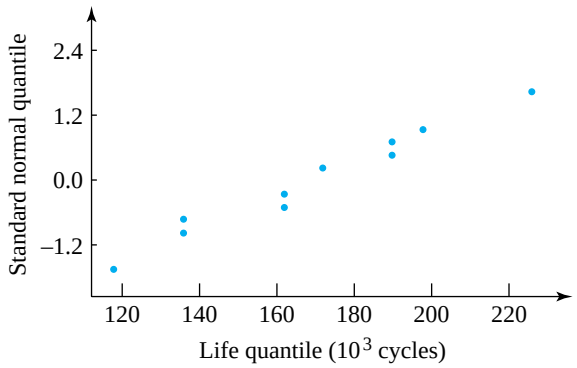

Figure 6.11 Normal plot of spring lifetimes

An important question here might be "What is the average spring lifetime under conditions of 950 N/mm^2 stress?" Since only $n = 10$ observations are available, the large-sample method of Section 6.1 is not applicable. Instead, only the method indicated by expression (6.20) is a possible option. For it to be appropriate, lifetimes must be normally distributed.

Without a relevant base of experience in materials, it is difficult to speculate a priori about the appropriateness of a normal lifetime model in this context. But at least it is possible to examine the data in Table 6.4 themselves for evidence of strong departure from normality. Figure 6.11 is a normal plot for the data. It shows that in fact no such evidence exists.

For the ten lifetimes, $\bar{x} = 168.3$ ($\times 10^3$ cycles) and $s = 33.1$ ($\times 10^3$ cycles). So to estimate the mean spring lifetime, these values may be used in expression (6.20), along with an appropriately chosen value of t. Using, for example, a 90% confidence level and a two-sided interval, t should be chosen as the .95 quantile of the t distribution with $\nu = n - 1 = 9$ degrees of freedom. That is, one uses the t_9 distribution and chooses $t > 0$ such that

$$P[-t < \text{a } t_9 \text{ random variable} < t] = .90$$

Consulting Table B.4, the choice $t = 1.833$ is in order. So a two-sided 90% confidence interval for μ has endpoints

$$168.3 \pm 1.833\frac{33.1}{\sqrt{10}}$$

i.e.,

$$168.3 \pm 19.2$$

i.e.,

$$149.1 \times 10^3 \text{ cycles} \quad \text{and} \quad 187.5 \times 10^3 \text{ cycles}$$

As illustrated in Example 8, normal-plotting the data as a rough check on the plausibility of an underlying normal distribution is a sound practice, and one that is used repeatedly in this text. However, it is important not to expect more than is justified from the method. It is certainly preferable to use it rather than making an unexamined leap to a possibly inappropriate normal assumption. But it is also true that when used with small samples, the method doesn't often provide definitive indications as to whether a normal model can be used. Small samples from normal distributions will often have only marginally linear-looking normal plots. At the same time, small samples from even quite nonnormal distributions can often have reasonably linear normal plots. In short, because of sampling variability, small samples don't carry much information about underlying distributional shape. About all that can be counted on from a small-sample preliminary normal plot, like that in Example 8, is a warning in case of gross departure from normality associated with an underlying distributional shape that is much heavier in the tails than a normal distribution (i.e., producing more extreme values than a normal shape would).

What is a "nonlinear" normal plot?

It is a good idea to make the effort to (so to speak) calibrate normal-plot perceptions if they are going to be used as a tool for checking a model. One way to do this is to use simulation and generate a number of samples of the size in question from a standard normal distribution and normal-plot these. Then the shape of the normal plot of the data in hand can be compared to the simulations to get some feeling as to whether any nonlinearity it exhibits is really unusual. To illustrate, Figure 6.12 shows normal plots for several simulated samples of size $n = 10$ from the standard normal distribution. Comparing Figures 6.11 and 6.12, it is clear that indeed the spring-life data carry no strong indication of nonnormality.

Example 8 shows the use of the confidence interval formula (6.20) but not the significance testing method (6.21). Since the small-sample method is exactly analogous to the large-sample method of Section 6.2 (except for the substitution of the t distribution for the standard normal distribution), and the source from which the data were taken doesn't indicate any particular value of μ belonging naturally in a null hypothesis, the use of the method indicated in expression (6.21) by itself will not be illustrated at this point. (There is, however, an application of the testing method to paired differences in Example 9.)

Small sample tests for μ

6.3.2 Inference for the Mean of Paired Differences

An important type of application of the foregoing methods of confidence interval estimation and significance testing is to *paired data*. In many engineering problems, it is natural to make two measurements of essentially the same kind, but differing in timing or physical location, on a single sample of physical objects. The goal in such situations is often to investigate the possibility of consistent differences between the two measurements. (Review the discussion of paired data terminology in Section 1.2.)

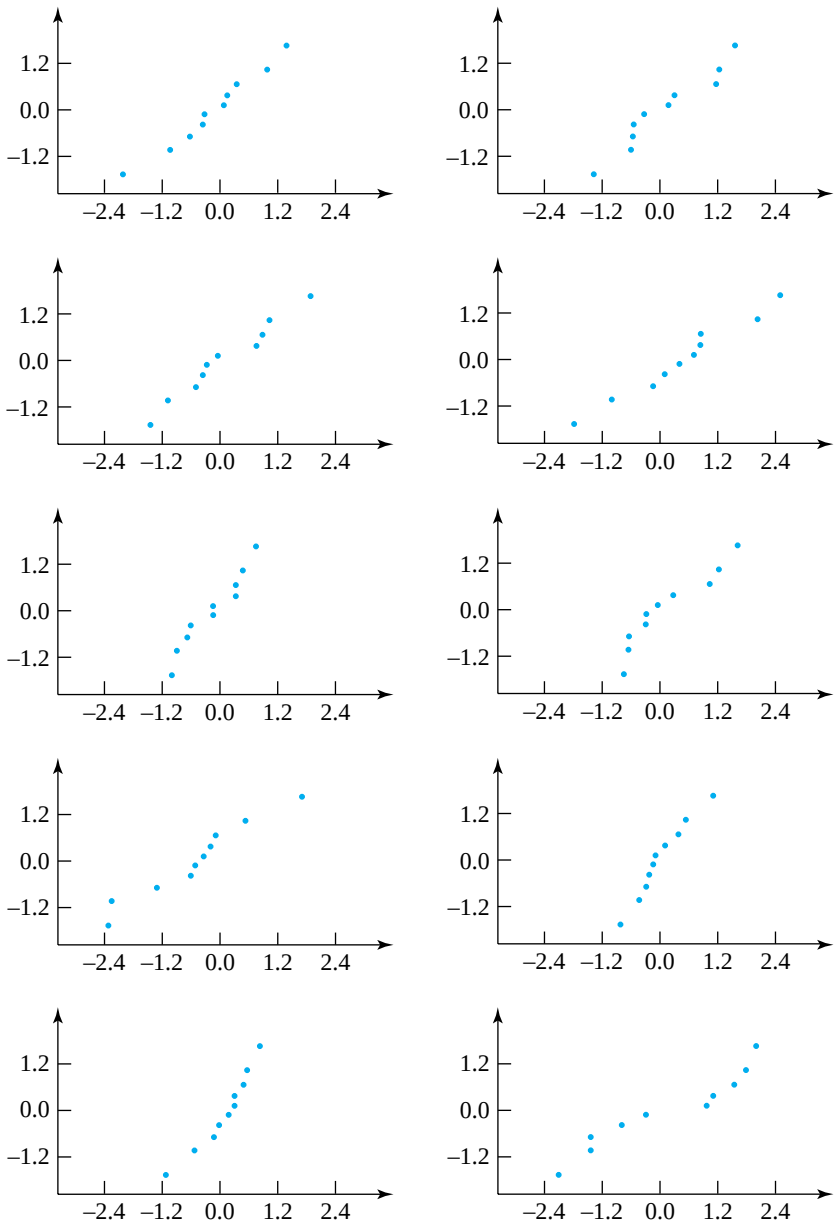

Figure 6.12 Normal plots of samples of size $n = 10$ from a standard normal distribution (data quantiles on the horizontal axes)

Example 9

Comparing Leading-Edge and Trailing-Edge Measurements on a Shaped Wood Product

Drake, Hones, and Mulholland worked with a company on the monitoring of the operation of an end-cut router in the manufacture of a wood product. They measured a critical dimension of a number of pieces of a particular type as they came off the router. Both a leading-edge and a trailing-edge measurement were made on each piece. The design for the piece in question specified that both leading-edge and trailing-edge values were to have a target value of .172 in. Table 6.5 gives leading- and trailing-edge measurements taken by the students on five consecutive pieces.

Table 6.5
Leading-Edge and Trailing-Edge Dimensions for Five Workpieces

Piece	Leading-Edge Measurement (in.)	Trailing-Edge Measurement (in.)
1	.168	.169
2	.170	.168
3	.165	.168
4	.165	.168
5	.170	.169

In this situation, the correspondence between leading- and trailing-edge dimensions was at least as critical to proper fit in a later assembly operation as was the conformance of the individual dimensions to the nominal value of .172 in. This was thus a paired-data situation, where one issue of concern was the possibility of a consistent difference between leading- and trailing-edge dimensions that might be traced to a machine misadjustment or unwise method of router operation.

In situations like Example 9, one simple method of investigating the possibility of a consistent difference between paired data is to first reduce the two measurements on each physical object to a single difference between them. Then the methods of confidence interval estimation and significance testing studied thus far may be applied to the differences. That is, after reducing paired data to differences $d_1, d_2, \ldots, d_n$, if n (the number of data pairs) is large, endpoints of a confidence interval for the underlying mean difference, μ_d, are

Large-sample confidence limits for μ_d

$$\bar{d} \pm z\frac{s_d}{\sqrt{n}} \tag{6.22}$$

where s_d is the sample standard deviation of $d_1, d_2, \ldots, d_n$. Similarly, the null hypothesis

$$H_0: \mu_d = \# \tag{6.23}$$

can be tested using the test statistic

Large-sample test statistic for μ_d

$$Z = \frac{\bar{d} - \#}{\frac{s_d}{\sqrt{n}}} \tag{6.24}$$

and a standard normal reference distribution.

If n is small, in order to come up with methods of formal inference, an underlying normal distribution *of differences* must be plausible. If that is the case, a confidence interval for μ_d has endpoints

Normal distribution confidence limits for μ_d

$$\bar{d} \pm t\frac{s_d}{\sqrt{n}} \tag{6.25}$$

and the null hypothesis (6.23) can be tested using the test statistic

Normal distribution test statistic for μ_d

$$T = \frac{\bar{d} - \#}{\frac{s_d}{\sqrt{n}}} \tag{6.26}$$

and a t_{n-1} reference distribution.

Example 9 (*continued*)

To illustrate this method of paired differences, consider testing the null hypothesis $H_0: \mu_d = 0$ and making a 95% confidence interval for any consistent difference between leading- and trailing-edge dimensions, μ_d, based on the data in Table 6.5.

Begin by reducing the $n = 5$ paired observations in Table 6.5 to differences

$$d = \text{leading-edge dimension} - \text{trailing-edge dimension}$$

appearing in Table 6.6. Figure 6.13 is a normal plot of the $n = 5$ differences in Table 6.6. A little experimenting with normal plots of simulated samples of size $n = 5$ from a normal distribution will convince you that the lack of linearity in Figure 6.13 would in no way be atypical of normal data. This, together with the fact that normal distributions are very often appropriate for describ-

Example 9
(*continued*)

Table 6.6
Five Differences in Leading- and Trailing-Edge Measurements

Piece	d = Difference in Dimensions (in.)	
1	−.001	(= .168 − .169)
2	.002	(= .170 − .168)
3	−.003	(= .165 − .168)
4	−.003	(= .165 − .168)
5	.001	(= .170 − .169)

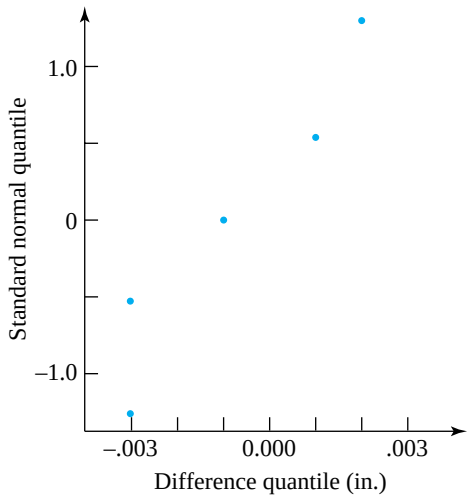

Figure 6.13 Normal plot of $n = 5$ differences

ing machined dimensions of mass-produced parts, suggests the conclusion that the methods represented by expressions (6.25) and (6.26) are in order in this example.

The differences in Table 6.6 have $\bar{d} = -.0008$ in. and $s_d = .0023$ in. So, first investigating the plausibility of a "no consistent difference" hypothesis in a five-step significance testing format, gives the following:

1. $H_0: \mu_d = 0$.
2. $H_a: \mu_d \neq 0$.
   (There is a priori no reason to adopt a one-sided alternative hypothesis.)

3. The test statistic will be

$$T = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

The reference distribution will be the t distribution with $\nu = n - 1 = 4$ degrees of freedom. Large observed $|t|$ will count as evidence against H_0 and in favor of H_a.

4. The sample gives

$$t = \frac{-.0008}{\frac{.0023}{\sqrt{5}}} = -.78$$

5. The observed level of significance is $P[|\text{a } t_4 \text{ random variable}| \geq .78]$, which can be seen from Table B.4 to be larger than $2(.10) = .2$. The data in hand are not convincing in favor of a systematic difference between leading- and trailing-edge measurements.

Consulting Table B.4 for the .975 quantile of the t_4 distribution, $t = 2.776$ is the appropriate multiplier for use in expression (6.25) for 95% confidence. That is, a two-sided 95% confidence interval for the mean difference between the leading- and trailing-edge dimensions has endpoints

$$-.0008 \pm 2.776\frac{.0023}{\sqrt{5}}$$

i.e.,

$$-.0008 \text{ in.} \pm .0029 \text{ in.} \qquad \textbf{(6.27)}$$

i.e.,

$$-.0037 \text{ in.} \quad \text{and} \quad .0021 \text{ in.}$$

This confidence interval for μ_d implicitly says (since 0 is in the calculated interval) that the observed level of significance for testing H_0: $\mu_d = 0$ is more than .05 ($= 1 - .95$). Put slightly differently, it is clear from display (6.27) that the imprecision represented by the plus-or-minus part of the expression is large enough to make it believable that the perceived difference, $\bar{d} = -.0008$, is just a result of sampling variability.

Large-sample inference for μ_d

Example 9 treats a small-sample problem. No example for large n is included here, because after the taking of differences just illustrated, such an example would reduce to a rehash of things in Sections 6.1 and 6.2. In fact, since for large n the t distribution with $\nu = n - 1$ degrees of freedom becomes essentially standard normal, one could even imitate Example 9 for large n and get into no logical problems. So at this point, it makes sense to move on from consideration of the paired-difference method.

6.3.3 Large-Sample Comparisons of Two Means (Based on Independent Samples)

One of the principles of effective engineering data collection discussed in Section 2.3 was *comparative study*. The idea of paired differences provides inference methods of a very special kind for comparison, where one sample of items in some sense provides its own basis for comparison. Methods that can be used to compare two means where two different "unrelated" samples form the basis of inference are studied next, beginning with large-sample methods.

Example 10

Comparing the Packing Properties of Molded and Crushed Pieces of a Solid

A company research effort involved finding a workable geometry for molded pieces of a solid. One comparison made was between the weight of molded pieces of a particular geometry, that could be poured into a standard container, and the weight of irregularly shaped pieces (obtained through crushing), that could be poured into the same container. A series of 24 attempts to pack both molded and crushed pieces of the solid produced the data (in grams) that are given in Figure 6.14 in the form of back-to-back stem-and-leaf diagrams.

Notice that although the same number of molded and crushed weights are represented in the figure, there are two distinctly different samples represented. This is in no way comparable to the paired-difference situation treated in Example 9, and a different method of statistical inference is appropriate.

In situations like Example 10, it is useful to adopt subscript notation for both the parameters and the statistics—for example, letting μ_1 and μ_2 stand for underlying distributional means corresponding to the first and second conditions and $\bar{x}_1$ and $\bar{x}_2$ stand for corresponding sample means. Now if the two data-generating mechanisms are conceptually essentially equivalent to sampling with replacement from two distributions, Section 5.5 says that $\bar{x}_1$ has mean μ_1 and variance σ_1^2/n_1, and $\bar{x}_2$ has mean μ_2 and variance σ_2^2/n_2.

The difference in sample means $\bar{x}_1 - \bar{x}_2$ is a natural statistic to use in comparing μ_1 and μ_2. Proposition 1 in Chapter 5 (see page 307) implies that if it is reasonable

Molded		Crushed
7.9	11	
4.5, 3.6, 1.2	12	
9.8, 8.9, 7.9, 7.1, 6.1, 5.7, 5.1	12	
2.3, 1.3, 0.0	13	
8.0, 7.0, 6.5, 6.3, 6.2	13	
2.2, 0.1	14	
	14	
2.1, 1.2, 0.2	15	
	15	
	16	1.8
	16	5.8, 9.6
	17	1.3, 2.0, 2.4, 3.3, 3.4, 3.7
	17	6.6, 9.8
	18	0.2, 0.9, 3.3, 3.8, 4.9
	18	5.5, 6.5, 7.1, 7.3, 9.1, 9.8
	19	0.0, 1.0
	19	

Figure 6.14 Back-to-back stem-and-leaf plots of packing weights for molded and crushed pieces

to think of the two samples as separately chosen/independent, the random variable has

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

and

$$\text{Var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

If, in addition, n_1 and n_2 are large (so that $\bar{x}_1$ and $\bar{x}_2$ are each approximately normal), $\bar{x}_1 - \bar{x}_2$ is approximately normal—i.e.,

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{6.28}$$

has an approximately standard normal probability distribution.

It is possible to begin with the fact that the variable (6.28) is approximately standard normal and end up with confidence interval and significance-testing methods for $\mu_1 - \mu_2$ by using logic exactly parallel to that in the "known-σ" parts of Sections 6.1 and 6.2. But practically, it is far more useful to begin instead with an expression that is free of the parameters σ_1 and σ_2. Happily, for large n_1 and n_2, not only is the variable (6.28) approximately standard normal but so is

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \tag{6.29}$$

Then the standard logic of Section 6.1 shows that a two-sided large-sample confidence interval for the difference $\mu_1 - \mu_2$ based on two independent samples has endpoints

Large-sample confidence limits for $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 \pm z\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.30}$$

where z is chosen such that the probability that the standard normal distribution assigns to the interval between $-z$ and z corresponds to the desired confidence. And the logic of Section 6.2 shows that under the same conditions,

$$H_0: \mu_1 - \mu_2 = \#$$

can be tested using the statistic

Large-sample test statistic for $\mu_1 - \mu_2$

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - \#}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \tag{6.31}$$

and a standard normal reference distribution.

Example 10 (*continued*)

In the molding problem, the crushed pieces were a priori expected to pack better than the molded pieces (that for other purposes are more convenient). Consider testing the statistical significance of the difference in mean weights and also making a 95% one-sided confidence interval for the difference (declaring that the crushed mean weight minus the molded mean weight is at least some number).

The sample sizes here ($n_1 = n_2 = 24$) are borderline for being called large. It would be preferable to have a few more observations of each type. Lacking them, we will go ahead and use the methods of expressions (6.30) and (6.31) but

remain properly cautious of the results should they in any way produce a "close call" in engineering or business terms.

Arbitrarily labeling "crushed" condition 1 and "molded" condition 2 and calculating from the data in Figure 6.14 that $\bar{x}_1 = 179.55$ g, $s_1 = 8.34$ g, $\bar{x}_2 = 132.97$ g, and $s_2 = 9.31$ g, the five-step testing format produces the following summary:

1. $H_0: \mu_1 - \mu_2 = 0$.
2. $H_a: \mu_1 - \mu_2 > 0$.
   (The research hypothesis here is that the crushed mean exceeds the molded mean so that the difference, taken in this order, is positive.)
3. The test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

   The reference distribution is standard normal, and large observed values z will constitute evidence against H_0 and in favor of H_a.
4. The samples give

$$z = \frac{179.55 - 132.97 - 0}{\sqrt{\frac{(8.34)^2}{24} + \frac{(9.31)^2}{24}}} = 18.3$$

5. The observed level of significance is $P[\text{a standard normal variable} \geq 18.3] \approx 0$. The data present overwhelming evidence that $\mu_1 - \mu_2 > 0$—i.e., that the mean packed weight of crushed pieces exceeds that of the molded pieces.

Then turning to a one-sided confidence interval for $\mu_1 - \mu_2$, note that only the lower endpoint given in display (6.30) will be used. So $z = 1.645$ will be appropriate. That is, with 95% confidence, we conclude that the difference in means (crushed minus molded) exceeds

$$(179.55 - 132.97) - 1.645\sqrt{\frac{(8.34)^2}{24} + \frac{(9.31)^2}{24}}$$

i.e., exceeds

$$46.58 - 4.20 = 42.38 \text{ g}$$

Example 10 (*continued*)

Or differently put, a 95% one-sided confidence interval for $\mu_1 - \mu_2$ is

$$(42.38, \infty)$$

Students are sometimes uneasy about the arbitrary choice involved in labeling the two conditions in a two-sample study. The fact is that either one can be used. As long as a given choice is followed through consistently, the real-world conclusions reached will be completely unaffected by the choice. In Example 10, if the molded condition is labeled number 1 and the crushed condition number 2, an appropriate one-sided confidence for the molded mean minus the crushed mean is

$$(-\infty, -42.38)$$

This has the same meaning in practical terms as the interval in the example.

The present methods apply where single measurements are made on each element of two different samples. This stands in contrast to problems of paired data (where there are bivariate observations on a single sample). In the woodworking case of Example 9, the data were paired because both leading-edge and trailing-edge measurements were made on each piece. If leading-edge measurements were taken from one group of items and trailing-edge measurements from another, a two-sample (not a paired difference) analysis would be in order.

6.3.4 Small-Sample Comparisons of Two Means (Based on Independent Samples from Normal Distributions)

The last inference methods presented in this section are those for the difference in two means in cases where at least one of n_1 and n_2 is small. All of the discussion for this problem is limited to cases where observations are normal. And in fact, the most straightforward methods are for cases where, in addition, the two underlying standard deviations are comparable. The discussion begins with these.

Graphical check on the plausibility of the model

A way of making at least a rough check on the plausibility of "normal distributions with a common variance" model assumptions in an application is to normal-plot two samples on the same set of axes, checking not only for approximate linearity but also for approximate equality of slope.

Example 8 (*continued*)

The data of W. Armstrong on spring lifetimes (appearing in the book by Cox and Oakes) not only concern spring longevity at a 950 N/mm^2 stress level but also longevity at a 900 N/mm^2 stress level. Table 6.7 repeats the 950 N/mm^2 data from before and gives the lifetimes of ten springs at the 900 N/mm^2 stress level as well.

Table 6.7
Spring Lifetimes under Two Different Levels of Stress (10^3 cycles)

950 N/mm^2 Stress	900 N/mm^2 Stress
225, 171, 198, 189, 189	216, 162, 153, 216, 225
135, 162, 135, 117, 162	216, 306, 225, 243, 189

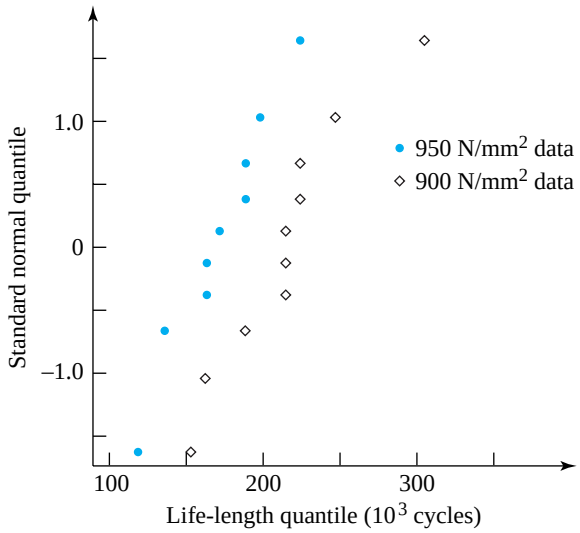

Figure 6.15 Normal plots of spring lifetimes under two different levels of stress

Figure 6.15 consists of normal plots for the two samples made on a single set of axes. In light of the kind of variation in linearity and slope exhibited in Figure 6.12 by the normal plots for samples of this size ($n = 10$) from a single normal distribution, there is certainly no strong evidence in Figure 6.15 against the appropriateness of an "equal variances, normal distributions" model for spring lifetimes.

If the assumption that $\sigma_1 = \sigma_2$ is used, then the common value is called σ, and it makes sense that both s_1 and s_2 will approximate σ. That suggests that they should somehow be combined into a single estimate of the basic, baseline variation. As it turns out, mathematical convenience dictates a particular method of combining or **pooling** the individual s's to arrive at a single estimate of σ.

Definition 14

If two numerical samples of respective sizes n_1 and n_2 produce respective sample variances s_1^2 and s_2^2, the **pooled sample variance**, s_p^2, is the weighted average of s_1^2 and s_2^2 where the weights are the sample sizes minus 1. That is,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad \textbf{(6.32)}$$

The **pooled sample standard deviation**, s_p, is the square root of s_p^2.

s_p is a kind of average of s_1 and s_2 that is guaranteed to fall between the two values s_1 and s_2. Its exact form is dictated more by considerations of mathematical convenience than by obvious intuition.

Example 8 (*continued*)

In the spring-life case, making the arbitrary choice to call the 900 N/mm^2 stress level condition 1 and the 950 N/mm^2 stress level condition 2, $s_1 = 42.9$ (10^3 cycles) and $s_2 = 33.1$ (10^3 cycles). So pooling the two sample variances via formula (6.32) produces

$$s_p^2 = \frac{(10 - 1)(42.9)^2 + (10 - 1)(33.1)^2}{(10 - 1) + (10 - 1)} = 1{,}468(10^3 \text{ cycles})^2$$

Then, taking the square root,

$$s_p = \sqrt{1{,}468} = 38.3(10^3 \text{ cycles})$$

In the argument leading to large-sample inference methods for $\mu_1 - \mu_2$, the quantity given in expression (6.28),

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

was briefly considered. In the $\sigma_1 = \sigma_2 = \sigma$ context, this can be rewritten as

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \textbf{(6.33)}$$

One could use the fact that expression (6.33) is standard normal to produce methods for confidence interval estimation and significance testing. But for use, these would require the input of the parameter σ. So instead of beginning with expression (6.28) or (6.33), it is standard to replace σ in expression (6.33) with s_{P} and begin with the quantity

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\text{P}}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{6.34}$$

Expression (6.34) is crafted exactly so that under the present model assumptions, the variable (6.34) has a well-known, tabled probability distribution: the t distribution with $\nu = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degrees of freedom. (Notice that the $n_1 - 1$ degrees of freedom associated with the first sample add together with the $n_2 - 1$ degrees of freedom associated with the second to produce $n_1 + n_2 - 2$ overall.) This probability fact, again via the kind of reasoning developed in Sections 6.1 and 6.2, produces inference methods for $\mu_1 - \mu_2$. That is, a two-sided confidence interval for the difference $\mu_1 - \mu_2$, based on independent samples from normal distributions with a common variance, has endpoints

Normal distributions ($\sigma_1 = \sigma_2$) confidence limits for $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 \pm t s_{\text{P}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{6.35}$$

where t is chosen such that the probability that the $t_{n_1+n_2-2}$ distribution assigns to the interval between $-t$ and t corresponds to the desired confidence. And under the same conditions,

$$\text{H}_0\colon \mu_1 - \mu_2 = \#$$

can be tested using the statistic

Normal distributions ($\sigma_1 = \sigma_2$) test statistic for $\mu_1 - \mu_2$

$$T = \frac{\bar{x}_1 - \bar{x}_2 - \#}{s_{\text{P}}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{6.36}$$

and a $t_{n_1+n_2-2}$ reference distribution.

Example 8 (*continued*)

We return to the spring-life case to illustrate small-sample inference for two means. First consider testing the hypothesis of equal mean lifetimes with an alternative of increased lifetime accompanying a reduction in stress level. Then

Example 8 (*continued*)

consider making a two-sided 95% confidence interval for the difference in mean lifetimes.

Continuing to call the 900 N/mm^2 stress level condition 1 and the 950 N/mm^2 stress level condition 2, from Table 6.7 $\bar{x}_1 = 215.1$ and $\bar{x}_2 = 168.3$, while (from before) $s_p = 38.3$. The five-step significance-testing format then gives the following:

1. H_0: $\mu_1 - \mu_2 = 0$.
2. H_a: $\mu_1 - \mu_2 > 0$.
   (The engineering expectation is that condition 1 produces the larger lifetimes.)
3. The test statistic is $T = \dfrac{\bar{x}_1 - \bar{x}_2 - 0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

   The reference distribution is t with $10 + 10 - 2 = 18$ degrees of freedom, and large observed t will count as evidence against H_0.
4. The samples give

$$t = \frac{215.1 - 168.3 - 0}{38.3\sqrt{\dfrac{1}{10} + \dfrac{1}{10}}} = 2.7$$

5. The observed level of significance is $P[\text{a } t_{18} \text{ random variable} \geq 2.7]$, which (according to Table B.4) is between .01 and .005. This is strong evidence that the lower stress level is associated with larger mean spring lifetimes.

Then, if the expression (6.35) is used to produce a two-sided 95% confidence interval, the choice of t as the .975 quantile of the t_{18} distribution is in order. Endpoints of the confidence interval for $\mu_1 - \mu_2$ are

$$(215.1 - 168.3) \pm 2.101(38.3)\sqrt{\frac{1}{10} + \frac{1}{10}}$$

i.e.,

$$46.8 \pm 36.0$$

i.e.,

$$10.8 \times 10^3 \text{ cycles} \quad \text{and} \quad 82.8 \times 10^3 \text{ cycles}$$

The data in Table 6.7 provide enough information to establish convincingly that increased stress is associated with reduced mean spring life. But although the apparent size of that reduction when moving from the 900 N/mm^2 level (condition 1) to the 950 N/mm^2 level (condition 2) is 46.8×10^3 cycles, the variability present in the data is large enough (and the sample sizes small enough) that only a precision of $\pm 36.0 \times 10^3$ cycles can be attached to the figure 46.8×10^3 cycles.

Small-sample inference for $\mu_1 - \mu_2$ without the $\sigma_1 = \sigma_2$ assumption

There is no completely satisfactory answer to the question of how to do inference for $\mu_1 - \mu_2$ when it is not sensible to assume that $\sigma_1 = \sigma_2$. The most widely accepted (but approximate) method for the problem is one due to Satterthwaite that is related to the large-sample formula (6.30). That is, while endpoints (6.30) are not appropriate when n_1 or n_2 is small (they don't produce actual confidence levels near the nominal one), a modification of them is appropriate. Let

Satterthwaite's "estimated degrees of freedom"

$$\hat{\nu} = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{s_1^4}{(n_1 - 1)n_1^2} + \dfrac{s_2^4}{(n_2 - 1)n_2^2}} \tag{6.37}$$

and for a desired confidence level, suppose that $\hat{t}$ is such that the t distribution with $\hat{\nu}$ degrees of freedom assigns that probability to the interval between $-\hat{t}$ and $\hat{t}$. Then the two endpoints

Satterthwaite (approximate) normal distribution confidence limits for $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 \pm \hat{t}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.38}$$

can serve as confidence limits for $\mu_1 - \mu_2$ with a confidence level approximating the desired one. (One of the two limits (6.38) may be used as a single confidence bound with the two-sided unconfidence level halved.)

Example 8 (*continued*)

Armstrong collected spring lifetime data at stress levels besides the 900 and 950 N/mm^2 levels used thus far in this example. Ten springs tested at 850 N/mm^2 had lifetimes with $\bar{x} = 348.1$ and $s = 57.9$ (both in 10^3 cycles) and a reasonably linear normal plot. But taking the 850, 900, and 950 N/mm^2 data together, there is a clear trend to smaller and more consistent lifetimes as stress is increased. In light of this fact, should mean lifetimes at the 850 and 950 N/mm^2 stress levels be compared, use of a constant variance assumption seems questionable.

Example 8 (*continued*)

Consider then what the Satterthwaite method (6.38) gives for two-sided approximate 95% confidence limits for the difference in 850 and 950 N/mm^2 mean lifetimes. Equation (6.37) gives

$$\hat{\nu} = \frac{\left(\dfrac{(57.9)^2}{10} + \dfrac{(33.1)^2}{10}\right)^2}{\dfrac{(57.9)^4}{9(100)} + \dfrac{(33.1)^4}{9(100)}} = 14.3$$

and (rounding "degrees of freedom" down) the .975 quantile of the t_{14} distribution is 2.145. So the 95% limits (6.38) for the (850 N/mm^2 minus 950 N/mm^2) difference in mean lifetimes ($\mu_{850} - \mu_{950}$) are

$$348.1 - 168.3 \pm 2.145\sqrt{\frac{(57.9)^2}{10} + \frac{(33.1)^2}{10}}$$

i.e.,

$$179.8 \pm 45.2$$

i.e.,

$$134.6 \times 10^3 \text{ cycles} \quad \text{and} \quad 225.0 \times 10^3 \text{ cycles}$$

The inference methods represented by displays (6.35), (6.36), and (6.38) are the last of the standard one- and two-sample methods for means. In the next two sections, parallel methods for variances and proportions are considered. But before leaving this section to consider those methods, a final comment is appropriate about the small-sample methods.

This discussion has emphasized that, strictly speaking, the nominal properties (in terms of coverage probabilities for confidence intervals and relevant p-value declarations for significance tests) of the small-sample methods depend on the appropriateness of exactly normal underlying distributions and (in the cases of the methods (6.35) and (6.36)) exactly equal variances. On the other hand, when actually applying the methods, rather crude probability-plotting checks have been used for verifying (only) that the models are roughly plausible. According to conventional statistical wisdom, the small-sample methods presented here are remarkably robust to all but gross departures from the model assumptions. That is, as long as the model assumptions are at least roughly a description of reality, the nominal confidence levels and p-values will not be ridiculously incorrect. (For example, a nominally 90% confidence interval method might in reality be only an 80% method, but it will not be only a 20% confidence interval method.) So the kind of plotting that has been illustrated here is often taken as adequate precaution against unjustified application of the small-sample inference methods for means.

Section 3 Exercises

1. What is the practical consequence of using a "normal distribution" confidence interval formula when in fact the underlying data-generating mechanism cannot be adequately described using a normal distribution? Say something more specific/informative than "an error might be made," or "the interval might not be valid." (What, for example, can be said about the real confidence level that ought to be associated with a nominally 90% confidence interval in such a situation?)

2. Consider again the situation of Exercise 3 of Section 3.1. (It concerns the torques required to loosen two particular bolts holding an assembly on a piece of machinery.)
(a) What model assumptions are needed in order to do inference for the mean top-bolt torque here? Make a plot to investigate the necessary distributional assumption.
(b) Assess the strength of the evidence in the data that the mean top-bolt torque differs from a target value of 100 ft lb.
(c) Make a two-sided 98% confidence interval for the mean top-bolt torque.
(d) What model assumptions are needed in order to compare top-bolt and bottom-bolt torques here? Make a plot for investigating the necessary distributional assumption.
(e) Assess the strength of the evidence that there is a mean increase in required torque as one moves from the top to the bottom bolts.
(f) Give a 98% two-sided confidence interval for the mean difference in torques between the top and bottom bolts.

3. The machine screw measurement study of DuToit, Hansen, and Osborne referred to in Exercise 4 of Section 6.1 involved measurement of diameters of each of 50 screws with both digital and vernier-scale calipers. For the student referred to in that exercise, the differences in measured diameters (digital minus vernier, with units of mm) had the following frequency distribution:

Difference	−.03	−.02	−.01	.00	.01	.02
Frequency	1	3	11	19	10	6

(a) Make a 90% two-sided confidence interval for the mean difference in digital and vernier readings for this student.
(b) Assess the strength of the evidence provided by these differences to the effect that there is a systematic difference in the readings produced by the two calipers (at least when employed by this student).
(c) Briefly discuss why your answers to parts (a) and (b) of this exercise are compatible. (Discuss how the outcome of part (b) could easily have been anticipated from the outcome of part (a).)

4. B. Choi tested the stopping properties of various bike tires on various surfaces. For one thing, he tested both treaded and smooth tires on dry concrete. The lengths of skid marks produced in his study under these two conditions were as follows (in cm).

Treaded	Smooth
365, 374, 376	341, 348, 349
391, 401, 402	355, 375, 391

(a) In order to make formal inferences about $\mu_{\text{Treaded}} - \mu_{\text{Smooth}}$ based on these data, what must you be willing to use for model assumptions? Make a plot to investigate the reasonableness of those assumptions.
(b) Proceed under the necessary model assumptions to assess the strength of Choi's evidence of a difference in mean skid lengths.
(c) Make a 95% two-sided confidence interval for $\mu_{\text{Treaded}} - \mu_{\text{Smooth}}$ assuming that treaded and smooth skid marks have the same variability.
(d) Use the Satterthwaite method and make an approximate 95% two-sided confidence interval for $\mu_{\text{Treaded}} - \mu_{\text{Smooth}}$ assuming only that skid mark lengths for both types of tires are normally distributed.

6.4 One- and Two-Sample Inference for Variances

This text has repeatedly indicated that engineers must often pay close attention to the measurement, the prediction, and sometimes the physical reduction of variability associated with a system response. Accordingly, it makes sense to consider inference for a single variance and inference for comparing two variances. In doing so, two more standard families of probability distributions—the χ^2 distributions and the F distributions—will be introduced.

6.4.1 Inference for the Variance of a Normal Distribution

The key step in developing most of the formal inference methods discussed in this chapter has been to find a random quantity involving both the parameter (or function of parameters) of interest and sample-based quantities that under appropriate assumptions can be shown to have some well-known distribution. Inference methods for a single variance rely on a type of continuous probability distribution that has not yet been discussed in this book: the χ^2 distributions.

Definition 15

The χ^2 **(Chi-squared) distribution with degrees of freedom parameter**, ν, is a continuous probability distribution with probability density

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma\left(\dfrac{\nu}{2}\right)} x^{(\nu/2)-1} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \textbf{(6.39)}$$

If a random variable has the probability density given by formula (6.39), it is said to have the χ^2_ν distribution.

Form (6.39) is not terribly inviting, but neither is it unmanageable. For instance, it is easy enough to use it to make the kind of plots in Figure 6.16 for comparing the shapes of the χ^2_ν distributions for various choices of ν.

The χ^2_ν distribution has mean ν and variance 2ν. For $\nu = 2$, it is exactly the exponential distribution with mean 2. For large ν, the χ^2_ν distributions look increasingly bell-shaped (and can in fact be approximated by normal distributions with matching means and variances). Rather than using form (6.39) to find χ^2 probabilities, it is more common to use tables of χ^2 quantiles. Table B.5 is one such table. Across the top of the table are several cumulative probabilities. Down the left side of the table are values of the degrees of freedom parameter, ν. In the body of the table are corresponding quantiles.

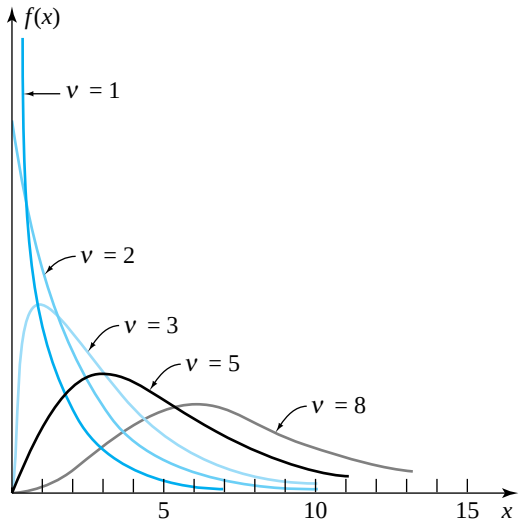

Figure 6.16 χ^2 probability densities for $\nu = 1, 2, 3, 5$, and 8

Example 11

Use of a Table of χ^2 Distribution Quantiles

Suppose that V is a random variable with a χ^2_3 distribution. Consider first finding the .95 quantile of V's distribution and then seeing what Table B.5 says about $P[V < .4]$ and $P[V > 10.0]$.

Using the χ^2 table, Table B.5

First, looking at the $\nu = 3$ row of Table B.5 under the cumulative probability .95, one finds 7.815 in the body of the table. That is, $Q(.95) = 7.815$, or (equivalently) $P[V \leq 7.815] = .95$. Then note that again using the $\nu = 3$ line of Table B.5, .4 lies between the .05 and .10 quantiles of the χ^2_3 distribution. Thus,

$$.05 < P[V < .4] < .10$$

Finally, since 10.0 lies between the ($\nu = 3$ line) entries of the table corresponding to cumulative probabilities .975 and .99 (i.e., the .975 and .99 quantiles of the χ^2_3 distribution), one may reason that

$$.01 < P[V > 10.0] < .025$$

The χ^2 distributions are of interest here because of a probability fact concerning the behavior of the random variable s^2 if the observations from which it is calculated are iid normal random variables. Under such assumptions,

$$X^2 = \frac{(n-1)s^2}{\sigma^2} \tag{6.40}$$

has a χ^2_{n-1} distribution. This fact is what is needed to identify inference methods for σ.

That is, given a desired confidence level concerning σ, one can choose χ^2 quantiles (say, L and U) such that the probability that a χ^2_{n-1} random variable will take a value between L and U corresponds to that confidence level. (Typically, L and U are chosen to "split the 'unconfidence' between the upper and lower χ^2_{n-1} tails"—for example, using the .05 and .95 χ^2_{n-1} quantiles for L and U, respectively, if 90% confidence is of interest.) Then, because the variable (6.40) has a χ^2_{n-1} distribution, the probability that

$$L < \frac{(n-1)s^2}{\sigma^2} < U \tag{6.41}$$

corresponds to the desired confidence level. But expression (6.41) is algebraically equivalent to the eventuality that

$$\frac{(n-1)s^2}{U} < \sigma^2 < \frac{(n-1)s^2}{L}$$

This then means that when an engineering data-generating mechanism can be thought of as essentially equivalent to random sampling from a normal distribution, a two-sided confidence interval for σ^2 has endpoints

Normal distribution confidence limits for σ^2

$$\boxed{\frac{(n-1)s^2}{U} \quad \text{and} \quad \frac{(n-1)s^2}{L}} \tag{6.42}$$

where L and U are such that the χ^2_{n-1} probability assigned to the interval (L, U) corresponds to the desired confidence.

Further, there is an obvious significance-testing method for σ^2. That is, subject to the same modeling limitations needed to support the confidence interval method,

$$\mathrm{H}_0\colon \sigma^2 = \#$$

can be tested using the statistic

Normal distribution test statistic for σ^2

$$\boxed{X^2 = \frac{(n-1)s^2}{\#}} \tag{6.43}$$

and a χ^2_{n-1} reference distribution.

p-values for testing $H_0\colon \sigma^2 = \#$

One feature of the testing methodology that needs comment concerns the computing of p-values in the case that the alternative hypothesis is of the form $\mathrm{H_a}$: $\sigma^2 \neq \#$. (p-values for the one-sided alternative hypotheses $\mathrm{H_a}$: $\sigma^2 < \#$ and $\mathrm{H_a}$: $\sigma^2 > \#$ are, respectively, the left and right χ^2_{n-1} tail areas beyond the observed value

of X^2.) The fact that the χ^2 distributions have no point of symmetry leaves some doubt for two-sided significance testing as to how an observed value of X^2 should be translated into a (two-sided) p-value. The convention that will be used here is as follows. If the observed value is larger than the χ^2_{n-1} median, the (two-sided) p-value will be twice the χ^2_{n-1} probability to the right of the observed value. If the observed value of X^2 is smaller than the χ^2_{n-1} median, the (two-sided) p-value will be twice the χ^2_{n-1} probability to the left of the observed value.

Confidence limits for functions of σ^2

Knowing that display (6.42) gives endpoints for a confidence interval for σ^2 also leads to confidence intervals for functions of σ^2. The square roots of the values in display (6.42) give endpoints for a confidence interval for the standard deviation, σ. And six times the square roots of the values in display (6.42) could be used as endpoints of a confidence interval for the "6σ" **capability of a process**.

Example 12

Inference for the Capability of a CNC Lathe

Cowan, Renk, Vander Leest, and Yakes worked with a manufacturer of high-precision metal parts on a project involving a computer numerically controlled (CNC) lathe. A critical dimension of one particular part produced on the lathe had engineering specifications of the form

$$\text{Nominal dimension} \pm .0020 \text{ in.}$$

An important practical issue in such situations is whether or not the machine is capable of meeting specifications of this type. One way of addressing this is to collect data and do inference for the intrinsic machine short-term variability, represented as a standard deviation. Table 6.8 gives values of the critical dimension measured on 20 parts machined on the lathe in question over a three-hour period. The units are .0001 in. over nominal.

Table 6.8
Measurements of a Dimension on 20 Parts Machined on a CNC Lathe

Measured Dimension (.0001 in. over nominal)	Frequency
8	1
9	1
10	10
11	4
12	3
13	1

Example 12
(*continued*)

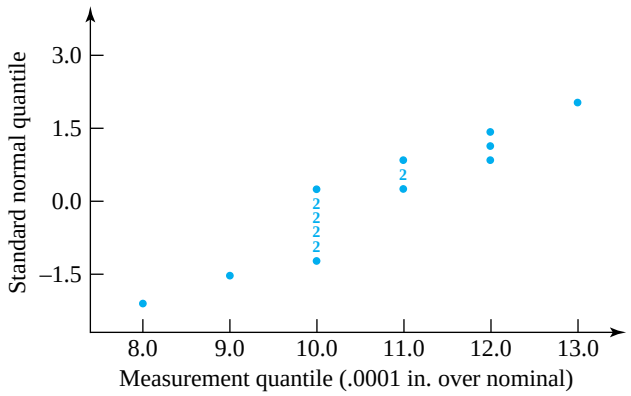

Figure 6.17 Normal plot of measurements on 20 parts machined on a CNC lathe

Suppose one takes the $\pm.0020$ in. engineering specifications as a statement of worst acceptable "$\pm 3\sigma$" machine capability, accordingly uses the data in Table 6.8, and (since $\frac{.0020}{3} \approx .0007$) tests $H_0: \sigma = .0007$. The relevance of the methods represented by displays (6.42) and (6.43) depends on the appropriateness of a normal distribution as a description of the critical dimension (as machined in the three-hour period in question). In this regard, note that (after allowing for the fact of the obvious discreteness of measurement introduced by gauging read to .0001 in.) the normal plot of the data from Table 6.8 shown in Figure 6.17 is not distressing in its departure from linearity. Further, at least over periods where manufacturing processes like the one in question are physically stable, normal distributions often prove to be quite adequate models for measured dimensions of mass-produced parts. Other evidence available on the machining process indicated that for practical purposes, the machining process was stable over the three-hour period in question. So one may proceed to use the normal-based methods, with no strong reason to doubt their relevance.

Direct calculation with the data of Table 6.8 shows that $s = 1.1 \times 10^{-4}$ in. So, using the five-step significance-testing format produces the following:

1. $H_0: \sigma = .0007$.
2. $H_a: \sigma > .0007$.
   (The most practical concern is the possibility that the machine is not capable of holding to the stated tolerances, and this is described in terms of σ larger than standard.)
3. The test statistic is

$$X^2 = \frac{(n-1)s^2}{(.0007)^2}$$

The reference distribution is χ^2 with $\nu = (20-1) = 19$ degrees of freedom, and large observed values x^2 (resulting from large values of s^2) will constitute evidence against H_0.

4. The sample gives

$$x^2 = \frac{(20-1)(.00011)^2}{(.0007)^2} = .5$$

5. The observed level of significance is $P[\text{a } \chi^2_{19} \text{ random variable} \geq .5]$. Now .5 is smaller than the .005 quantile of the χ^2_{19} distribution, so the p-value exceeds .995. There is nothing in the data in hand to indicate that the machine is incapable of holding to the given tolerances.

Consider, too, making a one-sided 99% confidence interval of the form $(0, \#)$ for 3σ. According to Table B.5, the .01 quantile of the χ^2_{19} distribution is $L = 7.633$. So using display (6.42), a 99% upper confidence bound for 3σ is

$$3\sqrt{\frac{(20-1)(1.1\times 10^{-4}\text{ in.})^2}{7.633}} = 5.0\times 10^{-4}\text{ in.}$$

When this is compared to the $\pm 20 \times 10^{-4}$ in. engineering requirement, it shows that the lathe in question is clearly capable of producing the kind of precision specified for the given dimension.

6.4.2 Inference for the Ratio of Two Variances (Based on Independent Samples from Normal Distributions)

To move from inference for a single variance to inference for comparing two variances requires the introduction of yet another new family of probability distributions: (Snedecor's) F distributions.

Definition 16

The **(Snedecor) F distribution with numerator and denominator degrees of freedom parameters ν_1 and ν_2** is a continuous probability distribution with probability density

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{\nu_1+\nu_2}{2}\right)\left(\dfrac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{(\nu_1/2)-1}}{\Gamma\left(\dfrac{\nu_1}{2}\right)\Gamma\left(\dfrac{\nu_2}{2}\right)\left(1+\dfrac{\nu_1 x}{\nu_2}\right)^{(\nu_1+\nu_2)/2}} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(6.44)}$$

If a random variable has the probability density given by formula (6.44), it is said to have the F_{ν_1,ν_2} distribution.

As Figure 6.18 reveals, the F distributions are strongly right-skewed distributions, whose densities achieve their maximum values at arguments somewhat less than 1. Roughly speaking, the smaller the values ν_1 and ν_2, the more asymmetric and spread out is the corresponding F distribution.

Direct use of formula (6.44) to find probabilities for the F distributions requires numerical integration methods. For purposes of applying the F distributions in statistical inference, the typical path is to instead make use of either statistical software or some fairly abbreviated tables of F distribution quantiles. Tables B.6 are tables of F quantiles. The body of a particular one of these tables, for a single p, gives the F distribution p quantiles for various combinations of ν_1 (the numerator degrees of freedom) and ν_2 (the denominator degrees of freedom). The values of ν_1 are given across the top margin of the table and the values of ν_2 down the left margin.

Using the F distribution tables, Tables B.6

Tables B.6 give only p quantiles for p larger than .5. Often F distribution quantiles for p smaller than .5 are needed as well. Rather than making up tables of such values, it is standard practice to instead make use of a computational trick. By using a relationship between F_{ν_1,ν_2} and F_{ν_2,ν_1} quantiles, quantiles for small p can be determined. If one lets Q_{ν_1,ν_2} stand for the F_{ν_1,ν_2} quantile function and Q_{ν_2,ν_1} stand for the quantile function for the F_{ν_2,ν_1} distribution,

Relationship between F_{ν_1,ν_2} and F_{ν_2,ν_1} quantiles

$$Q_{\nu_1,\nu_2}(p) = \frac{1}{Q_{\nu_2,\nu_1}(1-p)} \tag{6.45}$$

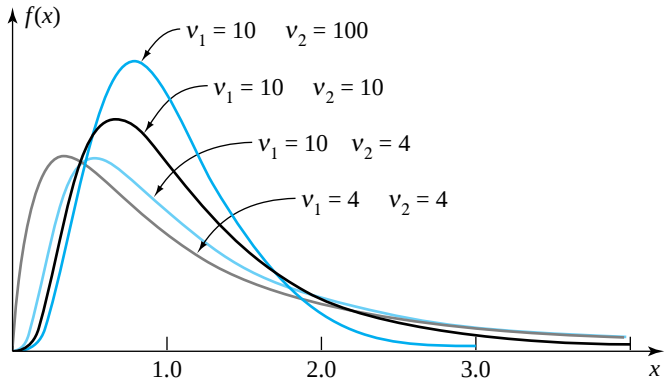

Figure 6.18 Four different F probability densities

Fact (6.45) means that a small lower percentage point of an F distribution may be obtained by taking the reciprocal of a corresponding small upper percentage point of the F distribution with degrees of freedom reversed.

Example 13

Use of Tables of *F* Distribution Quantiles

Suppose that V is an $F_{3,5}$ random variable. Consider finding the .95 and .01 quantiles of V's distribution and then seeing what Tables B.6 reveal about $P[V > 4.0]$ and $P[V < .3]$.

First, a direct look-up in the $p = .95$ table of quantiles, in the $\nu_1 = 3$ column and $\nu_2 = 5$ row, produces the number 5.41. That is, $Q(.95) = 5.41$, or (equivalently) $P[V < 5.41] = .95$.

To find the $p = .01$ quantile of the $F_{3,5}$ distribution, expression (6.45) must be used. That is,

$$Q_{3,5}(.01) = \frac{1}{Q_{5,3}(.99)}$$

so that using the $\nu_1 = 5$ column and $\nu_2 = 3$ row of the table of F .99 quantiles, one has

$$Q_{3,5}(.01) = \frac{1}{28.24} = .04$$

Next, considering $P[V > 4.0]$, one finds (using the $\nu_1 = 3$ columns and $\nu_2 = 5$ rows of Tables B.6) that 4.0 lies between the .90 and .95 quantiles of the $F_{3,5}$ distribution. That is,

$$.90 < P[V \leq 4.0] < .95$$

so that

$$.05 < P[V > 4.0] < .10$$

Finally, considering $P[V < .3]$, note that none of the entries in Tables B.6 is less than 1.00. So to place the value .3 in the $F_{3,5}$ distribution, one must locate its reciprocal, $3.33 (= 1/.3)$, in the $F_{5,3}$ distribution and then make use of expression (6.45). Using the $\nu_1 = 5$ columns and $\nu_2 = 3$ rows of Tables B.6, one finds that 3.33 is between the .75 and .90 quantiles of the $F_{5,3}$ distribution. So by expression (6.45), .3 is between the .1 and .25 quantiles of the $F_{3,5}$ distribution, and

$$.10 < P[V < .3] < .25$$

The extra effort required to find small F distribution quantiles is an artifact of standard table-making practice, rather than being any intrinsic extra difficulty associated with the F distributions. One way to eliminate the difficulty entirely is to use standard statistical software or a statistical calculator to find F quantiles.

The F distributions are of use here because a probability fact ties the behavior of ratios of independent sample variances based on samples from normal distributions to the variances σ_1^2 and σ_2^2 of those underlying distributions. That is, when s_1^2 and s_2^2 come from independent samples from normal distributions, the variable

$$F = \frac{s_1^2}{\sigma_1^2} \cdot \frac{\sigma_2^2}{s_2^2} \tag{6.46}$$

has an F_{n_1-1,n_2-1} distribution. (s_1^2 has $n_1 - 1$ associated degrees of freedom and is in the numerator of this expression, while s_2^2 has $n_2 - 1$ associated degrees of freedom and is in the denominator, providing motivation for the language introduced in Definition 16.)

This fact is exactly what is needed to produce formal inference methods for the ratio σ_1^2/σ_2^2. For example, it is possible to pick appropriate F quantiles L and U such that the probability that the variable (6.46) falls between L and U corresponds to a desired confidence level. (Typically, L and U are chosen to "split the 'unconfidence' " between the upper and lower F_{n_1-1,n_2-1} tails.) But

$$L < \frac{s_1^2}{\sigma_1^2} \cdot \frac{\sigma_2^2}{s_2^2} < U$$

is algebraically equivalent to

$$\frac{1}{U} \cdot \frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{L} \cdot \frac{s_1^2}{s_2^2}$$

That is, when a data-generating mechanism can be thought of as essentially equivalent to independent random sampling from two normal distributions, a two-sided confidence interval for σ_1^2/σ_2^2 has endpoints

Normal distributions confidence limits for σ_1^2/σ_2^2

$$\boxed{\frac{s_1^2}{U \cdot s_2^2} \quad \text{and} \quad \frac{s_1^2}{L \cdot s_2^2}} \tag{6.47}$$

where L and U are (F_{n_1-1,n_2-1} quantiles) such that the F_{n_1-1,n_2-1} probability assigned to the interval (L, U) corresponds to the desired confidence.

In addition, there is an obvious significance-testing method for σ_1^2/σ_2^2. That is, subject to the same modeling limitations as needed to support the confidence interval method,

$$\mathrm{H}_0\colon \frac{\sigma_1^2}{\sigma_2^2} = \# \tag{6.48}$$

can be tested using the statistic

Normal distributions test statistic for σ_1^2/σ_2^2

$$F = \frac{s_1^2/s_2^2}{\#} \tag{6.49}$$

and an F_{n_1-1,n_2-1} reference distribution. (The choice of $\# = 1$ in displays (6.48) and (6.49), so that the null hypothesis is one of equality of variances, is the only one commonly used in practice.) p-values for the one-sided alternative hypotheses $\mathrm{H_a}\colon \sigma_1^2/\sigma_2^2 < \#$ and $\mathrm{H_a}\colon \sigma_1^2/\sigma_2^2 > \#$ are (respectively) the left and right F_{n_1-1,n_2-1} tail areas beyond the observed values of the test statistic. For the two-sided alternative hypothesis $\mathrm{H_a}\colon \sigma_1^2/\sigma_2^2 \neq \#$, the standard convention is to report twice the F_{n_1-1,n_2-1} probability to the right of the observed f if $f > 1$ and to report twice the F_{n_1-1,n_2-1} probability to the left of the observed f if $f < 1$.

p-values for testing $H_0\colon \frac{\sigma_1^2}{\sigma_2^2} = \#$

Example 14

Comparing Uniformity of Hardness Test Results for Two Types of Steel

Condon, Smith, and Woodford did some hardness testing on specimens of 4% carbon steel. Part of their data are given in Table 6.9, where Rockwell hardness measurements for ten specimens from a lot of heat-treated steel specimens and five specimens from a lot of cold-rolled steel specimens are represented.

Consider comparing measured hardness *uniformity* for these two steel types (rather than mean hardness, as might have been done in Section 6.3). Figure 6.19 shows side-by-side dot diagrams for the two samples and suggests that there is a larger variability associated with the heat-treated specimens than with the cold-rolled specimens. The two normal plots in Figure 6.20 indicate no obvious problems with a model assumption of normal underlying distributions.

Table 6.9
Rockwell Hardness Measurements for Steel Specimens of Two Types

Heat-Treated	Cold-Rolled
32.8, 44.9, 34.4, 37.0, 23.6, 29.1, 39.5, 30.1, 29.2, 19.2	21.0, 24.5, 19.9, 14.8, 18.8

Example 14 (*continued*)

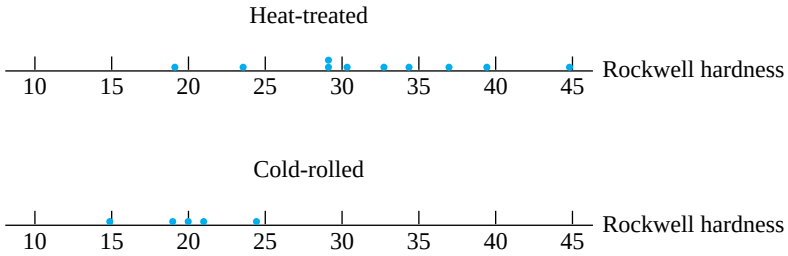

Figure 6.19 Dot diagrams of hardness for heat-treated and cold-rolled steels

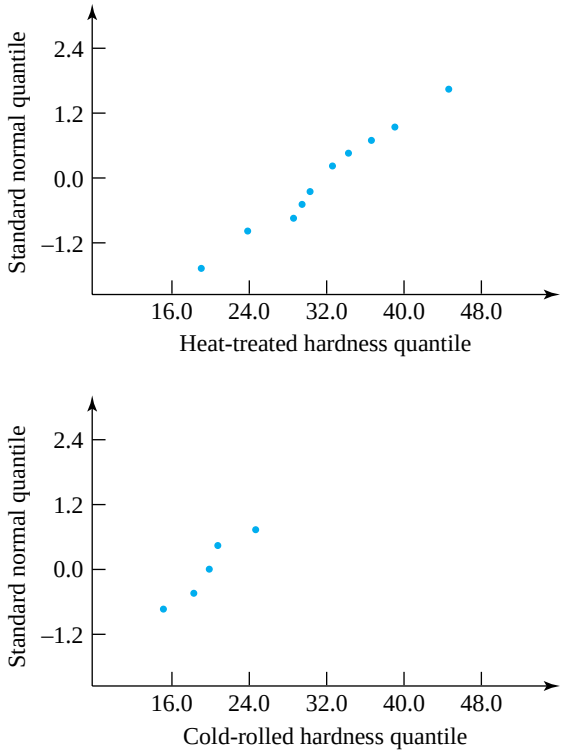

Figure 6.20 Normal plots of hardness for heat-treated and cold-rolled steels

Then, arbitrarily choosing to call the heat-treated condition number 1 and the cold-rolled condition 2, $s_1 = 7.52$ and $s_2 = 3.52$, and a five-step significance test of equality of variances based on the variable (6.49) proceeds as follows:

1. $H_0: \dfrac{\sigma_1^2}{\sigma_2^2} = 1.$

2. $H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$.
(If there is any materials-related reason to pick a one-sided alternative hypothesis here, the authors don't know it.)

3. The test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

The reference distribution is the $F_{9,4}$ distribution, and both large observed f and small observed f will constitute evidence against H_0.

4. The samples give

$$f = \frac{(7.52)^2}{(3.52)^2} = 4.6$$

5. Since the observed f is larger than 1, for the two-sided alternative, the p-value is

$$2P[\text{an } F_{9,4} \text{ random variable} \geq 4.6]$$

From Tables B.6, 4.6 is between the $F_{9,4}$ distribution .9 and .95 quantiles, so the observed level of significance is between .1 and .2. This makes it moderately (but not completely) implausible that the heat-treated and cold-rolled variabilities are the same.

In an effort to pin down the relative sizes of the heat-treated and cold-rolled hardness variabilities, the square roots of the expressions in display (6.47) may be used to give a 90% two-sided confidence interval for σ_1/σ_2. Now the .95 quantile of the $F_{9,4}$ distribution is 6.0, while the .95 quantile of the $F_{4,9}$ distribution is 3.63, implying that the .05 quantile of the $F_{9,4}$ distribution is $\frac{1}{3.63}$. Thus, a 90% confidence interval for the ratio of standard deviations σ_1/σ_2 has endpoints

$$\sqrt{\frac{(7.52)^2}{6.0(3.52)^2}} \quad \text{and} \quad \sqrt{\frac{(7.52)^2}{(1/3.63)(3.52)^2}}$$

That is,

.87 and 4.07

Example 14 (*continued*)

The fact that the interval (.87, 4.07) covers values both smaller and larger than 1 indicates that the data in hand do not provide definitive evidence even as to which of the two variabilities in material hardness is larger.

One of the most important engineering applications of the inference methods represented by displays (6.47) through (6.49) is in the comparison of inherent precisions for different pieces of equipment and for different methods of operating a single piece of equipment.

Example 15

Comparing Uniformities of Operation of Two Ream Cutters

Abassi, Afinson, Shezad, and Yeo worked with a company that cuts rolls of paper into sheets. The uniformity of the sheet lengths is important, because the better the uniformity, the closer the average sheet length can be set to the nominal value without producing undersized sheets, thereby reducing the company's giveaway costs. The students compared the uniformity of sheets cut on a ream cutter having a manual brake to the uniformity of sheets cut on a ream cutter that had an automatic brake. The basis of that comparison was estimated standard deviations of sheet lengths cut by the two machines—just the kind of information used to frame formal inferences in this section. The students estimated $\sigma_{\text{manual}}/\sigma_{\text{automatic}}$ to be on the order of 1.5 and predicted a period of two years or less for the recovery of the capital improvement cost of equipping all the company's ream cutters with automatic brakes.

The methods of this section are, strictly speaking, normal distribution methods. It is worthwhile to ask, "How essential is this normal distribution restriction to the predictable behavior of these inference methods for one and two variances?" There is a remark at the end of Section 6.3 to the effect that the methods presented there for *means* are fairly robust to moderate violation of the section's model assumptions. Unfortunately, such is *not* the case for the methods for *variances* presented here.

Caveats about inferences for variances

These are methods whose nominal confidence levels and p-values can be fairly badly misleading unless the normal models are good ones. This makes the kind of careful data scrutiny that has been implemented in the examples (in the form of normal-plotting) essential to the responsible use of the methods of this section. And it suggests that since normal-plotting itself isn't typically terribly revealing unless the sample size involved is moderate to large, formal inferences for variances will be most safely made on the basis of moderate to large normal-looking samples.

The importance of the "normal distribution(s)" restriction to the predictable operation of the methods of this section is not the only reason to prefer large sample sizes for inferences on variances. A little experience with the formulas in this section will convince the reader that (even granting the appropriateness of normal models) small samples often do not prove adequate to answer practical questions about variances. χ^2 and F confidence intervals for variances and variance ratios based on

small samples can be so big as to be of little practical value, and the engineer will typically be driven to large sample sizes in order to solve variance-related real-world problems. This is not in any way a failing of the present methods. It is simply a warning and quantification of the fact that learning about variances requires more data than (for example) learning about means.

Section 4 Exercises

1. Return to data on Choi's bicycle stopping distance given in Exercise 4 of Section 6.3.
   (a) Operating under the assumption that treaded tires produce normally distributed stopping distances, give a two-sided 95% confidence interval for the standard deviation of treaded tire stopping distances.
   (b) Operating under the assumption that smooth tires produce normally distributed stopping distances, give a 99% upper confidence bound for the standard deviation of smooth tire stopping distances.
   (c) Operating under the assumption that both treaded and smooth tires produce normally distributed stopping distances, assess the strength of Choi's evidence that treaded and smooth stopping distances differ in their variability. (Use H_0: $\sigma_{\text{Treaded}} = \sigma_{\text{Smooth}}$ and H_a: $\sigma_{\text{Treaded}} \neq \sigma_{\text{Smooth}}$ and show the whole five-step format.)
   (d) Operating under the assumption that both treaded and smooth tires produce normally distributed stopping distances, give a 90% two-sided confidence interval for the ratio $\sigma_{\text{Treaded}}/\sigma_{\text{Smooth}}$.
2. Consider again the situation of Exercise 3 of Section 3.1 and Exercise 2 of Section 6.3. (It concerns the torques required to loosen two particular bolts holding an assembly on a piece of machinery.)
   (a) Operating under the assumption that top-bolt torques are normally distributed, give a 95% lower confidence bound for the standard deviation of the top-bolt torques.
   (b) Translate your answer to part (a) into a 95% lower confidence bound on the "6σ process capability" of the top-bolt tightening process.
   (c) It is not appropriate to use the methods (6.47) through (6.49) and the data given in Exercise 3 of Section 3.1 to compare the consistency of top-bolt and bottom-bolt torques. Why?

6.5 One- and Two-Sample Inference for Proportions

The methods of formal statistical inference in the previous four sections are useful in the analysis of quantitative data. Occasionally, however, engineering studies produce only qualitative data, and one is faced with the problem of making properly hedged inferences from such data. This section considers how the sample fraction $\hat{p}$ (defined in Section 3.4) can be used as the basis for formal statistical inferences. It begins with the use of $\hat{p}$ from a single sample to make formal inferences about a single system or population. The section then treats the use of sample proportions from two samples to make inferences comparing two systems or populations.

6.5.1 Inference for a Single Proportion

Recall from display (3.6) (page 104) that the notation $\hat{p}$ is used for the fraction of a sample that possesses a characteristic of engineering interest. A sample of pellets produced by a pelletizing machine might prove individually conforming or nonconforming, and $\hat{p}$ could be the sample fraction conforming. Or in another case, a sample of turned steel shafts might individually prove acceptable, reworkable, or scrap; $\hat{p}$ could be the sample fraction reworkable.

If formal statistical inferences are to be based on $\hat{p}$, one must think of the physical situation in such a way that $\hat{p}$ is related to some parameter characterizing it. Accordingly, this section considers scenarios where $\hat{p}$ is derived from an *independent identical success/failure trials* data-generating mechanism. (See again Section 5.1.4 to review this terminology.) Applications will include inferences about physically stable processes, where p is a system's propensity to produce an item with the characteristic of interest. And they will include inferences drawn about population proportions p in enumerative contexts involving large populations. For example, the methods of this section can be used both to make inferences about the routine operation of a physically stable pelletizing machine and also to make inferences about the fraction of nonconforming machine parts contained in a specific lot of 10,000 such parts.

Review of the material on independent success/failure trials (and particularly the binomial distributions) in Section 5.1.4 should convince the reader that

$$X = n\hat{p} = \text{the number of items in the sample with the characteristic of interest}$$

has the binomial (n, p) distribution. The sample fraction $\hat{p}$ is just a scale change away from $X = n\hat{p}$, so facts about the distribution of X have immediate counterparts regarding the distribution of $\hat{p}$. For example, Section 5.1.4 stated that the mean and variance for the binomial (n, p) distribution are (respectively) np and $np(1-p)$. This (together with Proposition 1 in Chapter 5) implies that $\hat{p}$ has

Mean of the sample proportion

$$E\hat{p} = E\left(\frac{X}{n}\right) = \frac{1}{n}EX = \frac{1}{n}\cdot np = p \tag{6.50}$$

and

Variance of the sample proportion

$$\operatorname{Var}\hat{p} = \operatorname{Var}\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 \operatorname{Var} X = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \tag{6.51}$$

Equations (6.50) and (6.51) provide a reassuring picture of the behavior of the statistic $\hat{p}$. They show that the probability distribution of $\hat{p}$ is centered at the underlying parameter p, with a variability that decreases as n increases.

Example 16
(*Example 3, Chapter 5, revisited—page 234*)

Means and Standard Deviations of Sample Fractions of Reworkable Shafts

Return again to the case of the performance of a process for turning steel shafts. Assume for the time being that the process is physically stable and that the likelihood that a given shaft is reworkable is $p = .20$. Consider $\hat{p}$, the sample fraction of reworkable shafts in samples of first $n = 4$ and then $n = 100$ shafts.

Expressions (6.50) and (6.51) show that for the $n = 4$ sample size,

$$E\hat{p} = p = .2$$

$$\sqrt{\text{Var}\,\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(.2)(.8)}{4}} = .2$$

Similarly, for the $n = 100$ sample size,

$$E\hat{p} = p = .2$$

$$\sqrt{\text{Var}\,\hat{p}} = \sqrt{\frac{(.2)(.8)}{100}} = .04$$

Comparing the two standard deviations, it is clear that the effect of a change in sample size from $n = 4$ to $n = 100$ is to produce a factor of 5 ($= \sqrt{100/4}$) decrease in the standard deviation of $\hat{p}$, while the distribution of $\hat{p}$ is centered at p for both sample sizes.

Approximate normality of the sample proportion

The basic new insight needed to provide large-sample inference methods based on $\hat{p}$ is the fact that for large n, the binomial (n, p) distribution (and therefore also the distribution of $\hat{p}$) is approximately normal. That is, for large n, approximate probabilities for $X = n\hat{p}$ (or $\hat{p}$) can be found using the normal distribution with mean $\mu = np$ (or $\mu = p$) and variance $\sigma^2 = np(1-p)$ (or $\sigma^2 = \frac{p(1-p)}{n}$).

Example 16
(*continued*)

In the shaft-turning example, consider the probability that for a sample of $n = 100$ shafts, $\hat{p} \geq .25$. Notice that $\hat{p} \geq .25$ is equivalent here to the eventuality that $n\hat{p} \geq 25$. So in theory the form of the binomial probability function given in Definition 9 of Chapter 5 could be used and the desired probability could be evaluated exactly as

$$P[\hat{p} \geq .25] = P[X \geq 25] = f(25) + f(26) + \cdots + f(99) + f(100)$$

But instead of making such laborious calculations, it is common (and typically adequate for practical purposes) to settle instead for a normal approximation to probabilities such as this one.

Example 16 (*continued*)

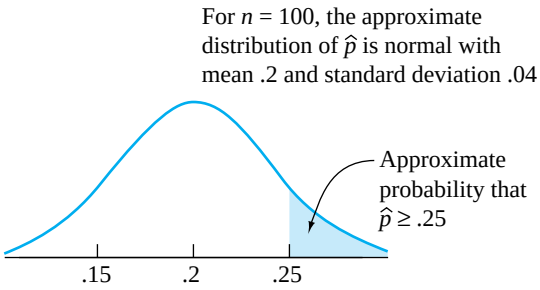

Figure 6.21 Approximate probability distribution for $\hat{p}$

Figure 6.21 shows the normal distribution with mean $\mu = p = .2$ and standard deviation $\sigma = \sqrt{p(1-p)/n} = .04$ and the corresponding probability assigned to the interval $[.25, \infty)$. Conversion of .25 to a z-value and then an approximate probability proceeds as follows:

$$z = \frac{.25 - E\hat{p}}{\sqrt{\text{Var}\,\hat{p}}} = \frac{.25 - .2}{.04} = 1.25$$

so

$$P[\hat{p} \geq .25] \approx 1 - \Phi(1.25) = .1056 \approx .11$$

The exact value of $P[\hat{p} \geq .25]$ (calculated to four decimal places using the binomial probability function) is .1314. (This can, for example, be obtained using the MINITAB routine under the "Calc/Probability Distributions/Binomial" menu.)

The statement that for large n, the random variable $\hat{p}$ is approximately normal is actually a version of the central limit theorem. For a given n, the approximation is best for moderate p (i.e., p near .5), and a common rule of thumb is to require that both the expected number of successes and the expected number of failures be at least 5 before making use of a normal approximation to the binomial (n, p) distribution. This is a requirement that

$$np \geq 5 \quad \text{and} \quad n(1-p) \geq 5$$

which amounts to a requirement that

Conditions for the normal approximation to the binomial

$$5 \leq np \leq n - 5 \tag{6.52}$$

(Notice that in Example 16, $np = 100(.2) = 20$ and $5 \leq 20 \leq 95$.)

An alternative, and typically somewhat stricter rule of thumb (which comes from a requirement that the mean of the binomial distribution be at least 3 standard deviations from both 0 and n) is to require that

Another set of conditions for the normal approximation to the binomial

$$9 \leq (n+9)p \leq n \tag{6.53}$$

before using the normal approximation. (Again in Example 16, $(n+9)p = (100+9)(.2) = 21.8$ and $9 \leq 21.8 \leq 100$.)

The approximate normality of $\hat{p}$ for large n implies that for large n,

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \tag{6.54}$$

is approximately standard normal. This and the reasoning of Section 6.2 then imply that the null hypothesis

$$H_0: p = \#$$

can be tested using the statistic

Large-sample test statistic for p

$$Z = \frac{\hat{p} - \#}{\sqrt{\dfrac{\#(1-\#)}{n}}} \tag{6.55}$$

and a standard normal reference distribution. Further, reasoning parallel to that in Section 6.1 (beginning with the fact that the variable (6.54) is approximately standard normal), leads to the conclusion that an interval with endpoints

$$\hat{p} \pm z\sqrt{\frac{p(1-p)}{n}} \tag{6.56}$$

(where z is chosen such that the standard normal probability between $-z$ and z corresponds to a desired confidence) is a mathematically valid two-sided confidence interval for p.

However, the endpoints indicated by expression (6.54) are of no practical use as they stand, since they involve the unknown parameter p. There are two standard ways of remedying this situation. One draws its motivation from the simple plot of $p(1-p)$ shown in Figure 6.22. That is, from Figure 6.22 it is easy to see that $p(1-p) \leq (.5)^2 = .25$, so the plus-or-minus part of formula (6.56) has (for $z > 0$)

$$z\sqrt{\frac{p(1-p)}{n}} \leq z\frac{1}{2\sqrt{n}}$$

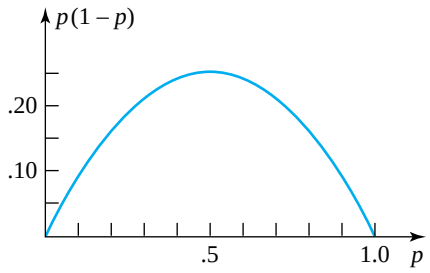

Figure 6.22 Plot of $p(1-p)$ versus p

Thus, modifying the endpoints in formula (6.56) by replacing the plus-or-minus part with $\pm z/2\sqrt{n}$ produces an interval that is guaranteed to be as wide as necessary to give the desired approximate confidence level. That is, the interval with endpoints

Large-sample conservative confidence limits for p

$$\hat{p} \pm z\frac{1}{2\sqrt{n}} \tag{6.57}$$

where z is chosen such that the standard normal probability between $-z$ and z corresponds to a desired confidence, is a practically usable large-n, two-sided, *conservative* confidence interval for p. (Appropriate use of only one of the endpoints in display (6.57) gives a one-sided confidence interval.)

The other common method of dealing with the fact that the endpoints in formula (6.56) are of no practical use is to begin the search for a formula from a point other than the approximate standard normal distribution of the variable (6.54). For large n, not only is the variable (6.54) approximately standard normal, but so is

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} \tag{6.58}$$

And the denominator of the quantity (6.58) (which amounts to an estimated standard deviation for $\hat{p}$) is free of the parameter p. So when manipulations parallel to those in Section 6.1 are applied to expression (6.58), the conclusion is that the interval with endpoints

Large-sample confidence limits for p

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{6.59}$$

can be used as a two-sided, large-n confidence interval for p with confidence level corresponding to the standard normal probability assigned to the interval between $-z$ and z. (One-sided confidence limits are obtained in the usual way, using only one of the endpoints in display (6.59) and appropriately adjusting the confidence level.)

Example 17

Inference for the Fraction of Dry Cells with Internal Shorts

The article "A Case Study of the Use of an Experimental Design in Preventing Shorts in Nickel-Cadmium Cells" by Ophir, El-Gad, and Snyder (*Journal of Quality Technology*, 1988) describes a series of experiments conducted to find how to reduce the proportion of cells scrapped by a battery plant because of internal shorts. At the beginning of the study, about 6% of the cells produced were being scrapped because of internal shorts.

Among a sample of 235 cells made under a particular trial set of plant operating conditions, 9 cells had shorts. Consider what formal inferences can be drawn about the set of operating conditions based on such data. $\hat{p} = \frac{9}{235} = .038$, so two-sided 95% confidence limits for p, are by expression (6.59)

$$.038 \pm 1.96\sqrt{\frac{(.038)(1-.038)}{235}}$$

i.e.,

$$.038 \pm .025$$

i.e.,

$$.013 \quad \text{and} \quad .063 \tag{6.60}$$

Notice that according to display (6.60), although $\hat{p} = .038 < .06$ (and thus indicates that the trial conditions were an improvement over the standard ones), the case for this is not airtight. The data in hand allow some possibility that p for the trial conditions even exceeds .06. And the ambiguity is further emphasized if the conservative formula (6.57) is used in place of expression (6.59). Instead of 95% confidence endpoints of $.038 \pm .025$, formula (6.57) gives endpoints $.038 \pm .064$.

To illustrate the significance-testing method represented by expression (6.55), consider testing with an alternative hypothesis that the trial plant conditions are an improvement over the standard ones. One then has the following summary:

1. H_0: $p = .06$.
2. H_a: $p < .06$.
3. The test statistic is

$$Z = \frac{\hat{p} - .06}{\sqrt{\frac{(.06)(1-.06)}{n}}}$$

The reference distribution is standard normal, and small observed values z will count as evidence against H_0.

Example 17 (*continued*)

4. The sample gives

$$z = \frac{.038 - .06}{\sqrt{\dfrac{(.06)(1 - .06)}{235}}} = -1.42$$

5. The observed level of significance is then

$$\Phi(-1.42) = .08$$

This is strong but not overwhelming evidence that the trial plant conditions are an improvement on the standard ones.

It needs to be emphasized again that these inferences depend for their practical relevance on the appropriateness of the "stable process/independent, identical trials" model for the battery-making process and extend only as far as that description continues to make sense. It is important that the experience reported in the article was gained under (presumably physically stable) regular production, so there is reason to hope that a single "independent, identical trials" model can describe both experimental and future process behavior.

Sample size determination for estimating p

Section 6.1 illustrated the fact that the form of the large-n confidence interval for a mean can be used to guide sample-size choices for estimating μ. The same is true regarding the estimation of p. If one (1) has in mind a desired confidence level, (2) plans to use expression (6.57) or has in mind a worst-case (largest) expectation for $\hat{p}(1 - \hat{p})$ in expression (6.59), and (3) has a desired precision of estimation of p, it is a simple matter to solve for a corresponding sample size. That is, suppose that the desired confidence level dictates the use of the value z in formula (6.57) and one wants to have confidence limits (or a limit) of the form $\hat{p} \pm \Delta$. Setting

$$\Delta = z\frac{1}{2\sqrt{n}}$$

and solving for n produces the requirement

$$n = \left(\frac{z}{2\Delta}\right)^2$$

Example 17 (*continued*)

Return to the nicad battery case and suppose that for some reason a better fix on the implications of the new operating conditions was desired. In fact, suppose that p is to be estimated with a two-sided conservative 95% confidence interval, and $\pm.01$ (fraction defective) precision of estimation is desired. Then, using the

plus-or-minus part of expression (6.57) (or equivalently, the plus-or-minus part of expression (6.59) under the worst-case scenario that $\hat{p} = .5$), one is led to set

$$.01 = 1.96\frac{1}{2\sqrt{n}}$$

From this, a sample size of

$$n \approx 9{,}604$$

is required.

In most engineering contexts this sample size is impractically large. Rethinking the calculation by planning the use of expression (6.59) and adopting the point of view that, say, 10% is a worst-case expectation for $\hat{p}$ (and thus $.1(1 - .1) = .09$ is a worst-case expectation for $\hat{p}(1 - \hat{p})$), one might be led instead to set

$$.01 = 1.96\sqrt{\frac{(.1)(1 - .1)}{n}}$$

However, solving for n, one has

$$n \approx 3{,}458$$

which is still beyond what is typically practical.

The moral of these calculations is that something has to give. The kind of large confidence and somewhat precise estimation requirements set at the beginning here cannot typically be simultaneously satisfied using a realistic sample size. One or the other of the requirements must be relaxed.

Cautions concerning inference based on sample proportions

The sample-size conclusions just illustrated are typical, and they justify two important points about the use of qualitative data. First, qualitative data carry less information than corresponding numbers of quantitative data (and therefore usually require very large samples to produce definitive inferences). This makes measurements generally preferable to qualitative observations in engineering applications. Second, if inferences about p based on even large values of n are often disappointing in their precision or reliability, there is little practical motivation to consider small-sample inference for p in a beginning text like this.

6.5.2 Inference for the Difference Between Two Proportions (Based on Independent Samples)

Two separately derived sample proportions $\hat{p}_1$ and $\hat{p}_2$, representing different processes or populations, can enable formal comparison of those processes or populations. The logic behind those methods of inference concerns the difference $\hat{p}_1 - \hat{p}_2$. If

1. the "independent, identical success-failure trials" description applies separately to the mechanisms that generate two samples,
2. the two samples are reasonably described as independent, and
3. both n_1 and n_2 are large,

a very simple approximate description of the distribution of $\hat{p}_1 - \hat{p}_2$ results.

Assuming $\hat{p}_1$ and $\hat{p}_2$ are independent, Proposition 1 in Chapter 5 and the discussion in this section concerning the mean and variance of a single sample proportion imply that $\hat{p}_1 - \hat{p}_2$ has

Mean of a difference in sample proportions

$$E(\hat{p}_1 - \hat{p}_2) = E\hat{p}_1 + (-1)E\hat{p}_2 = p_1 - p_2 \tag{6.61}$$

and

Variance of a difference in sample proportions

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = (1)^2 \text{Var}\,\hat{p}_1 + (-1)^2 \text{Var}\,\hat{p}_2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \tag{6.62}$$

Approximate normality of $\hat{p}_1 - \hat{p}_2$

Then the approximate normality of $\hat{p}_1$ and $\hat{p}_2$ for large sample sizes turns out to imply the approximate normality of the difference $\hat{p}_1 - \hat{p}_2$.

Example 16 (*continued*)

Consider again the turning of steel shafts, and imagine that two different, physically stable lathes produce reworkable shafts at respective rates of 20 and 25%. Then suppose that samples of (respectively) $n_1 = 50$ and $n_2 = 50$ shafts produced by the machines are taken, and the reworkable sample fractions $\hat{p}_1$ and $\hat{p}_2$ are found. Consider approximating the probability that $\hat{p}_1 \geq \hat{p}_2$ (i.e., that $\hat{p}_1 - \hat{p}_2 \geq 0$).

Using expressions (6.61) and (6.62), the variable $\hat{p}_1 - \hat{p}_2$ has

$$E(\hat{p}_1 - \hat{p}_2) = .20 - .25 = -.05$$

and

$$\sqrt{\text{Var}(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{(.20)(1-.20)}{50} + \frac{(.25)(1-.25)}{50}} = \sqrt{.00695} = .083$$

Figure 6.23 shows the approximately normal distribution of $\hat{p}_1 - \hat{p}_2$ and the area corresponding to $P[\hat{p}_1 - \hat{p}_2 \geq 0]$. The z-value corresponding to $\hat{p}_1 - \hat{p}_2 = 0$ is

$$z = \frac{0 - E(\hat{p}_1 - \hat{p}_2)}{\sqrt{\text{Var}(\hat{p}_1 - \hat{p}_2)}} = \frac{0 - (-.05)}{.083} = .60$$

so that

$$P[\hat{p}_1 - \hat{p}_2 \geq 0] = 1 - \Phi(.60) = .27$$

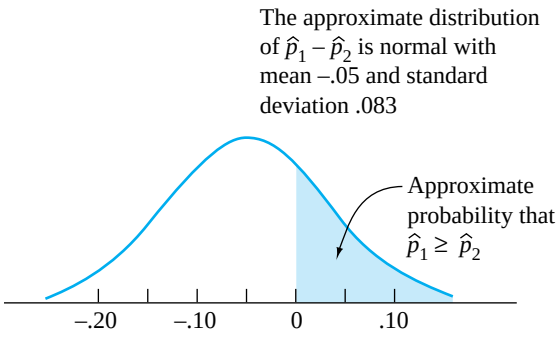

Figure 6.23 Approximate probability distribution for $\hat{p}_1 - \hat{p}_2$

The large-sample approximate normality of $\hat{p}_1 - \hat{p}_2$ translates to the realization that

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \tag{6.63}$$

is approximately standard normal, and this observation forms the basis for inference concerning $p_1 - p_2$. First consider confidence interval estimation for $p_1 - p_2$. The familiar argument of Section 6.1 (beginning with the quantity (6.63)) shows

$$\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{6.64}$$

to be a mathematically correct but practically unusable formula for endpoints of a confidence interval for $p_1 - p_2$. Conservative modification of expression (6.64), via replacement of both $p_1(1-p_1)$ and $p_2(1-p_2)$ with .25, shows that the two-sided interval with endpoints

Large-sample conservative confidence limits for $p_1 - p_2$

$$\hat{p}_1 - \hat{p}_2 \pm z \cdot \frac{1}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{6.65}$$

is a large-sample, two-sided, conservative confidence interval for $p_1 - p_2$ with confidence at least that corresponding to the standard normal probability between $-z$ and z. (One-sided intervals are obtained from expression (6.65) in the usual way.)

In addition, in by now familiar fashion, beginning with the fact that for large sample sizes, the modification of the variable (6.63),

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{6.66}$$

is approximately standard normal leads to the conclusion that the interval with endpoints

Large-sample confidence limits for $p_1 - p_2$

$$\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{6.67}$$

is a large-sample, two-sided confidence interval for $p_1 - p_2$ with confidence corresponding to the standard normal probability assigned to the interval between $-z$ and z. (Again, use of only one of the endpoints in display (6.67) gives a one-sided confidence interval.)

Example 18
(*Example 14, Chapter 3, revisited—page 111*)

Comparing Fractions Conforming for Two Methods of Operating a Pelletizing Process

Greiner, Grim, Larson, and Lukomski studied a number of different methods of running a pelletizing process. Two of these involved a mix with 20% reground powder with respectively small (condition 1) and large (condition 2) shot sizes. Of $n_1 = n_2 = 100$ pellets produced under these two sets of conditions, sample fractions $\hat{p}_1 = .38$ and $\hat{p}_2 = .29$ of the pellets conformed to specifications. Consider making a 90% confidence interval for comparing the two methods of process operation (i.e., an interval for $p_1 - p_2$).

Use of expression (6.67) shows that the interval with endpoints

$$.38 - .29 \pm 1.645\sqrt{\frac{(.38)(1-.38)}{100} + \frac{(.29)(1-.29)}{100}}$$

i.e.,

$$.09 \pm .109$$

i.e.,

$$-.019 \quad \text{and} \quad .199 \tag{6.68}$$

is a 90% confidence interval for $p_1 - p_2$, the difference in long-run fractions of conforming pellets that would be produced under the two sets of conditions. Notice that although appearances are that condition 1 has the higher associated likelihood of producing a conforming pellet, the case for this made by the data in hand is not airtight. The interval (6.68) allows some possibility that $p_1 - p_2 < 0$—i.e., that p_2 actually exceeds p_1. (The conservative interval indicated by expression (6.65) has endpoints of the form $.09 \pm .116$ and thus tells a similar story.)

The usual significance-testing method for $p_1 - p_2$ concerns the null hypothesis

$$H_0: p_1 - p_2 = 0 \tag{6.69}$$

i.e., the hypothesis that the parameters p_1 and p_2 are equal. Notice that if $p_1 = p_2$ and the common value is denoted as p, expression (6.63) can be rewritten as

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{6.70}$$

The variable (6.70) cannot serve as a test statistic for the null hypothesis (6.69), since it involves the unknown hypothesized common value of p_1 and p_2. What is done to modify the variable (6.70) to arrive at a usable test statistic, is to replace p with a sample-based estimate, obtained by pooling together the two samples. That is, let

Pooled estimator of a common p

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \tag{6.71}$$

($\hat{p}$ is the total number of items in the two samples with the characteristic of interest divided by the total number of items in the two samples). Then a significance test of hypothesis (6.69) can be carried out using the test statistic

Large-sample test statistic for $H_0: p_1 - p_2 = 0$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{6.72}$$

If H_0: $p_1 - p_2 = 0$ is true, Z in equation (6.72) is approximately standard normal, so a standard normal reference distribution is in order.

Example 18 (*continued*)

As further confirmation of the fact that in the pelletizing problem sample fractions of $\hat{p}_1 = .38$ and $\hat{p}_2 = .29$ based on samples of size $n_1 = n_2 = 100$ are not completely convincing evidence of a real difference in process performance for small and large shot sizes, consider testing H_0: $p_1 - p_2 = 0$ with H_a: $p_1 - p_2 \neq 0$. As a preliminary step, from expression (6.71),

$$\hat{p} = \frac{100(.38) + 100(.29)}{100 + 100} = \frac{67}{200} = .335$$

Then the five-step summary gives the following:

1. H_0: $p_1 - p_2 = 0$.
2. H_a: $p_1 - p_2 \neq 0$.
3. The test statistic is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

   The reference distribution is standard normal, and large observed values $|z|$ will constitute evidence against H_0.

4. The samples give

$$z = \frac{.38 - .29}{\sqrt{(.335)(1-.335)}\sqrt{\frac{1}{100} + \frac{1}{100}}} = 1.35$$

5. The p-value is $P[|\text{a standard normal variable}| \geq 1.35]$. That is, the p-value is

$$\Phi(-1.35) + \big(1 - \Phi(1.35)\big) = .18$$

   The data furnish only fairly weak evidence of a real difference in long-run fractions of conforming pellets for the two shot sizes.

The kind of results seen in Example 18 may take some getting used to. Even with sample sizes as large as 100, sample fractions differing by nearly .1 are still not necessarily conclusive evidence of a difference in p_1 and p_2. But this is just another manifestation of the point that *individual qualitative observations carry disappointingly little information*.

A final reminder of the large-sample nature of the methods presented here is in order. The methods here all rely (for the agreement of nominal and actual confidence

levels or the validity of their p-values) on the adequacy of normal approximations to binomial distributions. The approximations are workable provided expression (6.52) or (6.53) holds. When testing H_0: $p = \#$, it is easy to plug both n and # into expression (6.52) or (6.53) before putting great stock in normal-based p-values. But when estimating p or $p_1 - p_2$ or testing H_0: $p_1 - p_2 = 0$, no parallel check is obvious. So it is not completely clear how to screen potential applications for ones where the nominal confidence levels or p-values are possibly misleading. What is often done is to plug both n and $\hat{p}$ (or both n_1 and $\hat{p}_1$ and n_2 and $\hat{p}_2$) into expression (6.52) or (6.53) and verify that the inequalities hold before trusting nominal (normal-based) confidence levels and p-values. Since these random quantities are only approximations to the corresponding nonrandom quantities, one will occasionally be misled regarding the appropriateness of the normal approximations by such empirical checks. But they are better than automatic application, protected by no check at all.

Section 5 Exercises

1. Consider the situation of Example 14 of Chapter 3, and in particular the results for the 50% reground mixture.
   (a) Make and interpret 95% one-sided and two-sided confidence intervals for the fraction of conforming pellets that would be produced using the 50% mixture and the small shot size. (For the one-sided interval, give a lower confidence bound.) Use both methods of dealing with the fact that $\sigma_{\hat{p}}$ is not known and compare the resulting pairs of intervals.
   (b) If records show that past pelletizing performance was such that 55% of the pellets produced were conforming, does the value in Table 3.20 constitute strong evidence that the conditions of 50% reground mixture and small shot-size provide an improvement in yield? Show the five-step format.
   (c) Compare the small and large shot-size conditions using a 95% two-sided confidence interval for the difference in fractions conforming. Interpret the interval in the context of the example.
   (d) Assess the strength of the evidence given in Table 3.20 that the shot size affects the fraction of pellets conforming (when the 50% reground mixture is used).

2. In estimating a proportion p, a two-sided interval $\hat{p} \pm \Delta$ is used. Suppose that 95% confidence and $\Delta \leq .01$ are desired. About what sample size will be needed to guarantee this?

3. Specifications on the punch heights referred to in Chapter Exercise 9 of Chapter 3 were .500 in. to .505 in. In the sample of 405 punches measured by Hyde, Kuebrick, and Swanson, there were only 290 punches meeting these specifications. Suppose that the 405 punches can be thought of as a random sample of all such punches manufactured by the supplier under standard manufacturing conditions. Give an approximate 99% two-sided confidence interval for the standard fraction of nonconforming punches of this type produced by the punch supplier.

4. Consider two hypothetical machines producing a particular widget. If samples of $n_1 = 25$ and $n_2 = 25$ widgets produced by the respective machines have fractions nonconforming $\hat{p}_1 = .2$ and $\hat{p}_2 = .32$, is this strong evidence of a difference in machine nonconforming rates? What does this suggest about the kind of sample sizes typically needed in order to reach definitive conclusions based on attributes or qualitative data?

6.6 Prediction and Tolerance Intervals

Methods of confidence interval estimation and significance testing concern the problem of reasoning from sample information to statements about underlying *parameters* of the data generation, such as μ, σ, and p. These are extremely important engineering tools, but they often fail to directly address the question of real interest. Sometimes what is really needed as the ultimate product of a statistical analysis is not a statement about a parameter but rather an indication of reasonable bounds on other *individual values* generated by the process under study. For example, suppose you are about to purchase a new car. For some purposes, knowing that "the mean EPA mileage for this model is likely in the range 25 mpg $\pm$ .5 mpg" is not nearly as useful as knowing that "the EPA mileage figure for the particular car you are ordering is likely in the range 25 mpg $\pm$ 3 mpg." Both of these statements may be quite accurate, but they serve different purposes. The first statement is one about a *mean* mileage and the second is about an *individual* mileage. And it is only statements of the first type that have been directly treated thus far.

This section indicates what is possible in the way of formal statistical inferences, not for parameters but rather for individual values generated by a stable data-generating mechanism. There are two types of formal inference methods aimed in this general direction—statistical prediction interval methods and statistical tolerance interval methods—and both types will be discussed. The section begins with prediction intervals for a normal distribution. Then tolerance intervals for a normal distribution are considered. Finally, there is a discussion of how it is possible to use minimum and/or maximum values in a sample to create prediction and tolerance intervals for even nonnormal underlying distributions.

6.6.1 Prediction Intervals for a Normal Distribution

One fruitful way to phrase the question of inference for additional individual values produced by a process is the following: How might data in hand, $x_1, x_2, \ldots, x_n$, be used to create a numerical interval likely to bracket **one additional (as yet unobserved) value**, x_{n+1}, from the same data-generating mechanism? How, for example, might mileage tests on ten cars of a particular model be used to predict the results of the same test applied to an eleventh?

If the underlying distribution is adequately described as normal with mean μ and variance σ^2, there is a simple line of reasoning based on the random variable

$$\bar{x} - x_{n+1} \tag{6.73}$$

that leads to an answer to this question. That is, the random variable in expression (6.73) has, by the methods of Section 5.5 (Proposition 1 in particular),

$$E(\bar{x} - x_{n+1}) = E\bar{x} + (-1)Ex_{n+1} = \mu - \mu = 0 \tag{6.74}$$

and

$$\operatorname{Var}(\bar{x} - x_{n+1}) = (1)^2 \operatorname{Var}\bar{x} + (-1)^2 \operatorname{Var} x_{n+1} = \frac{\sigma^2}{n} + \sigma^2 = \left(1 + \frac{1}{n}\right)\sigma^2 \quad \textbf{(6.75)}$$

Further, it turns out that the difference (6.73) is normally distributed, so the variable

$$Z = \frac{(\bar{x} - x_{n+1}) - 0}{\sigma\sqrt{1 + \frac{1}{n}}} \quad \textbf{(6.76)}$$

is standard normal. And taking one more step, if s^2 is the usual sample variance of $x_1, x_2, \ldots, x_n$, substituting s for σ in expression (6.76) produces a variable

$$T = \frac{(\bar{x} - x_{n+1}) - 0}{s\sqrt{1 + \frac{1}{n}}} \quad \textbf{(6.77)}$$

which has a t distribution with $\nu = n - 1$ degrees of freedom.

Now (upon identifying x_{n+1} with μ and $\sqrt{1 + (1/n)}$ with $\sqrt{1/n}$), the variable (6.77) is formally similar to the t-distributed variable used to derive a small-sample confidence interval for μ. In fact, algebraic steps parallel to those used in the first part of Section 6.3 show that if $t > 0$ is such that the t_{n-1} distribution assigns, say, .95 probability to the interval between $-t$ and t, there is then .95 probability that

$$\bar{x} - ts\sqrt{1 + \frac{1}{n}} < x_{n+1} < \bar{x} + ts\sqrt{1 + \frac{1}{n}}$$

This reasoning suggests in general that the interval with endpoints

Normal distribution prediction limits for a single additional observation

$$\bar{x} \pm ts\sqrt{1 + \frac{1}{n}} \quad \textbf{(6.78)}$$

can be used as a two-sided interval to predict x_{n+1} and that the probability-based reliability figure attached to the interval should be the t_{n-1} probability assigned to the interval from $-t$ to t. The interval (6.78) is a called a **prediction interval** with associated confidence the t_{n-1} probability assigned to the interval from $-t$ to t. In general, the language indicated in Definition 17 will be used.

Definition 17

A **prediction interval** for a single additional observation is a data-based interval of numbers thought likely to contain the observation, possessing a stated probability-based confidence or reliability.

It is the fact that a finite sample gives only a somewhat clouded picture of a distribution that prevents the making of a normal distribution prediction interval from being a trivial matter of probability calculations like those in Section 5.2. That is, suppose there were enough data to "know" the mean, μ, and variance, σ^2, of a normal distribution. Then, since 1.96 is the .975 standard normal quantile, the interval with endpoints

$$\mu - 1.96\sigma \quad \text{and} \quad \mu + 1.96\sigma \tag{6.79}$$

has a 95% chance of bracketing the next value generated by the distribution. The fact that (when based only on small samples), the knowledge of μ and σ is noisy forces expression (6.79) to be abandoned for an interval like (6.78). It is thus comforting that for large n and 95% confidence, formula (6.78) produces an interval with endpoints approximating those in display (6.79). That is, for large n and 95% confidence, $t \approx 1.96$, $\sqrt{1 + (1/n)} \approx 1$, and one expects that typically $\bar{x} \approx \mu$ and $s \approx \sigma$, so that expressions (6.78) and (6.79) will essentially agree. The beauty of expression (6.78) is that it allows in a rational fashion for the uncertainties involved in the $\mu \approx \bar{x}$ and $\sigma \approx s$ approximations.

Example 19
(*Example 8 revisited*)

Predicting a Spring Lifetime

Recall from Section 6.3 that $n = 10$ spring lifetimes under 950 N/mm^2 stress conditions given in Table 6.4 (page 366) produced a fairly linear normal plot, $\bar{x} = 168.3$ ($\times 10^3$ cycles) and $s = 33.1$ ($\times 10^3$ cycles). Consider now predicting the lifetime of an additional spring of this type (under the same test conditions) with 90% confidence.

Using $\nu = 10 - 1 = 9$ degrees of freedom, the .95 quantile of the t distribution is (from Table B.4) 1.833. So, employing expression (6.78), there are two-sided 90% prediction limits for an additional spring lifetime

$$168.3 \pm 1.833(33.1)\sqrt{1 + \frac{1}{10}}$$

i.e.,

$$104.7 \times 10^3 \text{ cycles} \quad \text{and} \quad 231.9 \times 10^3 \text{ cycles} \tag{6.80}$$

The interval indicated by display (6.80) is not at all the same as the confidence interval for μ found in Example 8. The limits of

$$149.1 \times 10^3 \text{ cycles} \quad \text{and} \quad 187.5 \times 10^3 \text{ cycles}$$

found on page 367 apply to the mean spring lifetime, μ, not to an additional observation x_{11} as the ones in display (6.80) do.

Example 20

Predicting the Weight of a Newly Minted Penny

The delightful book *Experimentation and Measurement* by W. J. Youden (published as NBS Special Publication 672 by the U.S. Department of Commerce) contains a data set giving the weights of $n = 100$ newly minted U.S. pennies measured to 10^{-4} g but reported only to the nearest .02 g. These data are reproduced in Table 6.10. Figure 6.24 is a normal plot of these data and shows that a normal distribution is a plausible model for weights of newly minted pennies.

Further, calculation with the values in Table 6.10 shows that for the penny weights, $\bar{x} = 3.108$ g and $s = .043$ g. Then interpolation in Table B.4 shows the .9 quantile of the t_{99} distribution to be about 1.290, so that using only the "plus" part of expression (6.78), a one-sided 90% prediction interval of the form $(-\infty, \#)$ for the weight of a single additional penny has upper endpoint

$$3.108 + 1.290(.043)\sqrt{1 + \frac{1}{100}}$$

i.e.,

$$3.164 \text{ g} \tag{6.81}$$

Table 6.10
Weights of 100 Newly Minted U.S. Pennies

Penny Weight (g)	Frequency	Penny Weight (g)	Frequency
2.99	1	3.11	24
3.01	4	3.13	17
3.03	4	3.15	13
3.05	4	3.17	6
3.07	7	3.19	2
3.09	17	3.21	1

Example 20 (*continued*)

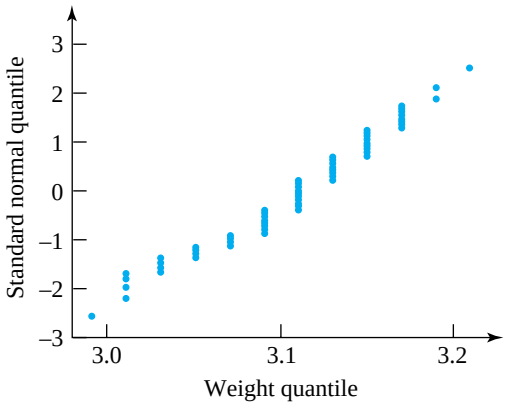

Figure 6.24 Normal plot of the penny weights

This example illustrates at least two important points. First, the two-sided prediction limits in display (6.78) can be modified to get a one-sided limit exactly as two-sided confidence limits can be modified to get a one-sided limit. Second, the calculation represented by the result (6.81) is, because $n = 100$ is a fairly large sample size, only marginally different from what one would get assuming $\mu = 3.108$ g exactly and $\sigma = .043$ g exactly. That is, since the .9 normal quantile is 1.282, "knowing" μ and σ leads to an upper prediction limit of

$$\mu + 1.282\sigma = 3.108 + (1.282)(.043) = 3.163 \text{ g} \qquad \textbf{(6.82)}$$

The fact that the result (6.81) is slightly larger than the final result in display (6.82) reflects the small uncertainty involved in the use of $\bar{x}$ in place of μ and s in place of σ.

Cautions about "prediction"

The name "prediction interval" probably has some suggested meanings that should be dismissed before going any further. *Prediction* suggests the future and thus potentially different conditions. But no such meaning should be associated with statistical prediction intervals. The assumption behind formula (6.78) is that $x_1, x_2, \ldots, x_n$ and x_{n+1} are *all* generated according to the *same* underlying distribution. If (for example, because of potential physical changes in a system during a time lapse between the generation of $x_1, x_2, \ldots, x_n$ and the generation of x_{n+1}) no single stable process model for the generation of all $n + 1$ observations is appropriate, then neither is formula (6.78). Statistical inference is not a crystal ball for foretelling an erratic and patternless future. It is rather a methodology for quantifying the extent of knowledge about a pattern of variation existing in a consistent present. It has implications in other times and at other places only if that same pattern of variation can be expected to repeat itself in those conditions.

It is also appropriate to comment on the meaning of the confidence or reliability figure attached to a prediction interval. Since a prediction interval is doing a different job than the confidence intervals of previous sections, the meaning of *confidence* given in Definition 2 doesn't quite apply here.

Prior to the generation of any of $x_1, x_2, \ldots, x_n, x_{n+1}$, planned use of expression (6.78) gives a guaranteed probability of success in bracketing x_{n+1}. And after all of $x_1, x_2, \ldots, x_n, x_{n+1}$ have been generated, one has either been completely successful or completely unsuccessful in bracketing x_{n+1}. But it is not altogether obvious how to think about "confidence" of prediction when $x_1, x_2, \ldots, x_n$ are in hand, but prior to the generation of x_{n+1}. For example, in the context of Example 19, having used sample data to arrive at the prediction limits in display (6.80)—i.e.,

$$104.7 \times 10^3 \text{ cycles} \quad \text{to} \quad 231.9 \times 10^3 \text{ cycles}$$

since x_{11} is a random variable, it would make sense to contemplate

$$P[104.7 \times 10^3 \leq x_{11} \leq 231.9 \times 10^3]$$

However, there is no guarantee on this probability nor any way to determine it. In particular, it is *not* necessarily .9 (the confidence level associated with the prediction interval). That is, there is no practical way to employ probability to describe the likely effectiveness of a numerical prediction interval. One is thus left with the interpretation of confidence of prediction given in Definition 18.

Definition 18
(*Interpretation of a Prediction Interval*)

To say that a numerical interval (a, b) is (for example) a 90% prediction interval for an additional observation x_{n+1} is to say that in obtaining it, methods of data collection and calculation have been applied that would produce intervals bracketing an $(n+1)$th observation in about 90% of repeated applications of the entire process of (1) selecting the sample $x_1, \ldots, x_n$, (2) calculating an interval, and (3) generating a single additional observation x_{n+1}. Whether or not x_{n+1} will fall into the numerical interval (a, b) is not known, and although there is some probability associated with that eventuality, it is not possible to evaluate it. And in particular, it need not be 90%.

When using a 90% prediction interval method, although some samples $x_1, \ldots, x_n$ produce numerical intervals with probability less than .9 of bracketing x_{n+1} and others produce numerical intervals with probability more than .9, the average for all samples $x_1, \ldots, x_n$ does turn out to be .9. The practical problem is simply that with data $x_1, \ldots, x_n$ in hand, you don't know whether you are above, below, or at the .9 figure.

6.6.2 Tolerance Intervals for a Normal Distribution

The emphasis, when making a prediction interval of the type just discussed, is on a *single* additional observation beyond those n already in hand. But in some practical engineering problems, many additional items are of interest. In such cases, one may wish to declare a data-based interval likely to encompass **most measurements from the rest of these items**.

Prediction intervals are not designed for the purpose of encompassing most of the measurements from the additional items of interest. The paragraph following Definition 18 argues that only on average is the fraction of a normal distribution bracketed by a 90% prediction interval equal to 90%. So a crude analysis (identifying the mean fraction bracketed with the median fraction bracketed) then suggests that the probability that the actual fraction bracketed is at least 90% is only about .5. That is, a 90% prediction interval is not constructed to be big enough for the present purpose. What is needed instead is a **statistical tolerance interval**.

Definition 19

A **statistical tolerance interval for a fraction p of an underlying distribution** is a data-based interval thought likely to contain at least a fraction p and possessing a stated (usually large) probability-based confidence or reliability.

The derivation of normal distribution tolerance interval formulas requires probability background well beyond what has been developed in this text. But results of that work look about as would be expected. It is possible, for a desired confidence level and fraction p of an underlying normal distribution, to find a corresponding constant τ_2 such that the two-sided interval with endpoints

Two-sided normal distribution tolerance limits

$$\bar{x} \pm \tau_2 s \tag{6.83}$$

is a tolerance interval for a fraction p of the underlying distribution. The τ_2 appearing in expression (6.83) is, for common (large) confidence levels, larger than the multiplier $t\sqrt{1+(1/n)}$ appearing in expression (6.78) for two-sided confidence of prediction p. On the other hand, as n gets large, both τ_2 from expression (6.83) and $t\sqrt{1+(1/n)}$ from expression (6.78) tend to the $(\frac{1+p}{2})$ standard normal quantile. Table B.7A gives some values of τ_2 for 95% and 99% confidence and $p = .9$, .95, and .99. (The use of this table will be demonstrated shortly.)

The factors τ_2 are not used to make one-sided tolerance intervals. Instead, another set of constants that will here be called τ_1 values have been developed. They are such that for a given confidence and fraction p of an underlying normal distribution, both of the one-sided intervals

A one-sided normal tolerance interval

$$(-\infty, \bar{x} + \tau_1 s) \tag{6.84}$$

and

Another one-sided normal tolerance interval

$$(\bar{x} - \tau_1 s, \infty) \tag{6.85}$$

are tolerance intervals for a fraction p of the distribution. τ_1 appearing in intervals (6.84) and (6.85) is, for common confidence levels, larger than the multiplier $t\sqrt{1 + (1/n)}$ appearing in expression (6.78) for one-sided confidence of prediction p. And as n gets large, both τ_1 from expression (6.84) or (6.85) and $t\sqrt{1 + (1/n)}$ from expression (6.78) tend to the standard normal p quantile. Table B.7B gives some values of τ_1.

Example 19 (*continued*)

Consider making a two-sided 95% tolerance interval for 90% of additional spring lifetimes based on the data of Table 6.4. As earlier, for these data, $\bar{x} = 168.3$ ($\times 10^3$ cycles) and $s = 33.1$ ($\times 10^3$ cycles). Then consulting Table B.7A, since $n = 10$, $\tau_2 = 2.856$ is appropriate for use in expression (6.83). That is, two-sided 95% tolerance limits for 90% of additional spring lifetimes are

$$168.3 \pm 2.856\,(33.1)$$

i.e.,

$$73.8 \times 10^3 \text{ cycles} \quad \text{and} \quad 262.8 \times 10^3 \text{ cycles} \tag{6.86}$$

It is obvious from comparing displays (6.80) and (6.86) that the effect of moving from the prediction of a single additional spring lifetime to attempting to bracket most of a large number of additional lifetimes is to increase the size of the declared interval.

Example 20 (*continued*)

Consider again the new penny weights given in Table 6.10 and now the problem of making a one-sided 95% tolerance interval of the form $(-\infty, \#)$ for the weights of 90% of additional pennies. Remembering that for the penny weights, $\bar{x} = 3.108$ g and $s = .043$ g, and using Table B.7B for $n = 100$, the desired upper tolerance bound for 90% of the penny weights is

$$3.108 + 1.527(.043) = 3.174 \text{ g}$$

As expected, this is larger (more conservative) than the value of 3.164 g given in display (6.81) as a one-sided 90% prediction limit for a single additional penny weight.

The correct interpretation of the confidence level for a tolerance interval should be fairly easy to grasp. Prior to the generation of $x_1, x_2, \ldots, x_n$, planned use of expression (6.83), (6.84), or (6.85) gives a guaranteed probability of success in bracketing a fraction of at least p of the underlying distribution. But after observing $x_1, \ldots, x_n$ and making a numerical interval, it is impossible to know whether the attempt has or has not been successful. Thus the following interpretation:

Definition 20 (*Interpretation of a Tolerance Interval*)

To say that a numerical interval (a, b) is (for example) a 90% tolerance interval for a fraction p of an underlying distribution is to say that in obtaining it, methods of data collection and calculation have been applied that would produce intervals bracketing a fraction of at least p of the underlying distribution in about 90% of repeated applications (of generation of $x_1, \ldots, x_n$ and subsequent calculation). Whether or not the numerical interval (a, b) actually contains at least a fraction p is unknown and not describable in terms of a probability.

6.6.3 Prediction and Tolerance Intervals Based on Minimum and/or Maximum Values in a Sample

Formulas (6.78), (6.83), (6.84), and (6.85) for prediction and tolerance limits are definitely normal distribution formulas. So what if an engineering data-generation process is stable but does not produce normally distributed observations? How, if at all, can prediction or tolerance limits be made? Two kinds of answers to these questions will be illustrated in this text. The first employs the transformation idea presented in Section 4.4, and the second involves the use of minimum and/or maximum sample values to establish prediction and/or tolerance bounds.

First (as observed in Section 4.4) if a response variable y fails to be normally distributed, it may still be possible to find some transformation g (essentially specifying a revised scale of measurement) such that $g(y)$ is normal. Then normal-based methods might be applied to $g(y)$ and answers of interest translated back into statements about y.

Example 21 (*Example 11, Chapter 4, revisited—page 192*)

Prediction and Tolerance Intervals for Discovery Times Obtained Using a Transformation

Section 5.3 argued that the auto service discovery time data of Elliot, Kibby, and Meyer given in Figure 4.31 (see page 192) are not themselves normal-looking, but that their natural logarithms are. This, together with the facts that the $n = 30$ natural logarithms have $\bar{x} = 2.46$ and $s = .68$, can be used to make prediction or tolerance intervals for log discovery times.

For example, using expression (6.78) to make a two-sided 99% prediction interval for an additional log discovery time produces endpoints

$$2.46 \pm 2.756(.68)\sqrt{1 + \frac{1}{30}}$$

i.e.,

$$.55 \ln \text{min} \quad \text{and} \quad 4.37 \ln \text{min} \tag{6.87}$$

And using expression (6.83) to make, for example, a 95% tolerance interval for 99% of additional log discovery times produces endpoints

$$2.46 \pm 3.355(.68)$$

i.e.,

$$.18 \ln \text{min} \quad \text{and} \quad 4.74 \ln \text{min} \tag{6.88}$$

Then the intervals specified in displays (6.87) and (6.88) for log discovery times have, via exponentiation, their counterparts for raw discovery times. That is, exponentiation of the values in display (6.87) gives a 99% prediction interval for another discovery time of from

$$1.7 \text{ min} \quad \text{to} \quad 79.0 \text{ min}$$

And exponentiation of the values in display (6.88) gives a 95% tolerance interval for 99% of additional discovery times of from

$$1.2 \text{ min} \quad \text{to} \quad 114.4 \text{ min}$$

When it is not possible to find a transformation that will allow normal-based methods to be used, prediction and tolerance interval formulas derived for other standard families of distributions (e.g., the Weibull family) can sometimes be appropriate. (The book *Statistical Intervals: A Guide for Practitioners*, by Hahn and Meeker, is a good place to look for these methods.) What can be done here is to point out that intervals from the smallest observation and/or to the largest value in a sample can be used as prediction and/or tolerance intervals for *any* underlying continuous distribution.

That is, if $x_1, x_2, \ldots, x_n$ are values in a sample and $\min(x_1, \ldots, x_n)$ and $\max(x_1, \ldots, x_n)$ are (respectively) the smallest and largest values among x_1, $x_2, \ldots, x_n$, consider the use of the intervals

Interval based on the sample maximum

$$(-\infty, \max(x_1, \ldots, x_n)) \tag{6.89}$$

and

Interval based on the sample minimum

$$(\min(x_1, \ldots, x_n), \infty) \tag{6.90}$$

and

Interval based on the sample minimum and maximum

$$(\min(x_1, \ldots, x_n), \max(x_1, \ldots, x_n)) \tag{6.91}$$

as prediction or tolerance intervals. Independent of exactly what underlying continuous distribution is operating, if the generation of $x_1, x_2, \ldots, x_n$ (and if relevant, x_{n+1}) can be described as a stable process, it is possible to evaluate the confidence levels associated with intervals (6.89), (6.90), and (6.91).

Consider first intervals (6.89) or (6.90) used as one-sided prediction intervals for a single additional observation x_{n+1}. The associated confidence level is

Prediction confidence for a one-sided interval

$$\text{One-sided prediction confidence level} = \frac{n}{n+1} \tag{6.92}$$

Then, considering interval (6.91) as a two-sided prediction interval for a single additional observation x_{n+1}, the associated confidence level is

Prediction confidence for a two-sided interval

$$\text{Two-sided prediction confidence level} = \frac{n-1}{n+1} \tag{6.93}$$

The confidence levels for intervals (6.89), (6.90), and (6.91) as tolerance intervals must of necessity involve p, the fraction of the underlying distribution one hopes to bracket. The fact is that using interval (6.89) or (6.90) as a one-sided tolerance interval for a fraction p of an underlying distribution, the associated confidence level is

Confidence level for a one-sided tolerance interval

$$\text{One-sided confidence level} = 1 - p^n \tag{6.94}$$

And when interval (6.91) is used as a tolerance interval for a fraction p of an underlying distribution, the appropriate associated confidence is

Confidence level for a two-sided tolerance interval

$$\text{Two-sided confidence level} = 1 - p^n - n(1-p)p^{n-1} \tag{6.95}$$

Example 19 (*continued*)

Return one more time to the spring-life scenario, and consider the use of interval (6.91) as first a prediction interval and then a tolerance interval for 90% of additional spring lifetimes. Notice in Table 6.4 (page 366) that the smallest and largest of the observed spring lifetimes are, respectively,

$$\min(x_1, \ldots, x_{10}) = 117 \times 10^3 \text{ cycles}$$

and

$$\max(x_1, \ldots, x_{10}) = 225 \times 10^3 \text{ cycles}$$

so the numerical interval under consideration is the one with endpoints 117 ($\times 10^3$ cycles) and 225 ($\times 10^3$ cycles).

Then expression (6.93) means that this interval can be used as a prediction interval with

$$\text{Prediction confidence} = \frac{10-1}{10+1} = \frac{9}{11} = 82\%$$

And expression (6.95) says that as a tolerance interval for a fraction $p = .9$ of many additional spring lifetimes, the interval can be used with associated confidence

$$\text{Confidence} = 1 - (.9)^{10} - 10(1-.9)(.9)^9 = 26\%$$

Example 20 (*continued*)

Looking for a final time at the penny weight data in Table 6.10, consider the use of interval (6.89) as first a prediction interval and then a tolerance interval for 99% of additional penny weights. Notice that in Table 6.10, the largest of the $n = 100$ weights is 3.21 g, so

$$\max(x_1, \ldots, x_{100}) = 3.21 \text{ g}$$

Example 20 (*continued*)

Then expression (6.92) says that when used as an upper prediction limit for a single additional penny weight, the prediction confidence associated with 3.21 g is

$$\text{Prediction confidence} = \frac{100}{100+1} = 99\%$$

And expression (6.94) shows that as a tolerance interval for 99% of many additional penny weights, the interval $(-\infty, 3.21)$ has associated confidence

$$\text{Confidence} = 1 - (.99)^{100} = 63\%$$

A little experience with formulas (6.92), (6.93), (6.94), and (6.95) will convince the reader that the intervals (6.89), (6.90), and (6.91) often carry disappointingly small confidence coefficients. Usually (but not always), you can do better in terms of high confidence and short intervals if (possibly after transformation) the normal distribution methods discussed earlier can be applied. But the beauty of intervals (6.89), (6.90), and (6.91) is that they are both widely applicable (in even nonnormal contexts) and extremely simple.

Prediction and tolerance interval methods are very useful engineering tools. Historically, they probably haven't been used as much as they should be for lack of accessible textbook material on the methods. We hope the reader is now aware of the existence of the methods as the appropriate form of formal inference when the focus is on individual values generated by a process rather than on process parameters. When the few particular methods discussed here don't prove adequate for practical purposes, the reader should look into the topic further, beginning with the book by Hahn and Meeker mentioned earlier.

Section 6 Exercises

1. Confidence, prediction, and tolerance intervals are all intended to do different jobs. What are these jobs? Consider the differing situations of an official of the EPA, a consumer about to purchase a single car, and a design engineer trying to equip a certain model with a gas tank large enough that most cars produced will have highway cruising ranges of at least 350 miles. Argue that depending on the point of view adopted, a lower confidence bound for a mean mileage, a lower prediction bound for an individual mileage, or a lower tolerance bound for most mileages would be of interest.
2. The 900 N/mm^2 stress spring lifetime data in Table 6.7 used in Example 8 have a fairly linear normal plot.
   (a) Make a two-sided 90% prediction interval for an additional spring lifetime under this stress.
   (b) Make a two-sided 95% tolerance interval for 90% of all spring lifetimes under this stress.
   (c) How do the intervals from (a) and (b) compare? (Consider both size and interpretation.)
   (d) There is a two-sided 90% confidence interval for the mean spring lifetime under this stress given in Example 8. How do your intervals from (a) and (b) compare to the interval in Example 8? (Consider both size and interpretation.)
   (e) Make a 90% lower prediction bound for an additional spring lifetime under this stress.

(f) Make a 95% lower tolerance bound for 90% of all spring lifetimes under this stress.

3. The natural logarithms of the aluminum contents discussed in Exercise 2 of Chapter 3 have a reasonably bell-shaped relative frequency distribution. Further, these 26 log aluminum contents have sample mean 4.9 and sample standard deviation .59. Use this information to respond to the following:
   (a) Give a two-sided 99% tolerance interval for 90% of additional log aluminum contents at the Rutgers recycling facility. Then translate this interval into a 99% tolerance interval for 90% of additional raw aluminum contents.
   (b) Make a 90% prediction interval for one additional log aluminum content and translate it into a prediction interval for a single additional aluminum content.
   (c) How do the intervals from (a) and (b) compare?
4. Again in the context of Chapter Exercise 2 of Chapter 3, if the interval from 30 ppm to 511 ppm is used as a prediction interval for a single additional aluminum content measurement from the study period, what associated prediction confidence level can be stated? What confidence can be associated with this interval as a tolerance interval for 90% of all such aluminum content measurements?

Chapter 6 Exercises

1. Consider the breaking strength data of Table 3.6. Notice that the normal plot of these data given as Figure 3.18 is reasonably linear. It may thus be sensible to suppose that breaking strengths for generic towel of this type (as measured by the students) are adequately modeled as normal. Under this assumption,
   (a) Make and interpret 95% two-sided and one-sided confidence intervals for the mean breaking strength of generic towels (make a one-sided interval of the form $(\#, \infty)$).
   (b) Make and interpret 95% two-sided and one-sided prediction intervals for a single additional generic towel breaking strength (for the one-sided interval, give the lower prediction bound).
   (c) Make and interpret 95% two-sided and one-sided tolerance intervals for 99% of generic towel breaking strengths (for the one-sided interval, give the lower tolerance bound).
   (d) Make and interpret 95% two-sided and one-sided confidence intervals for σ, the standard deviation of generic towel breaking strengths.
   (e) Put yourself in the position of a quality control inspector, concerned that the mean breaking strength not fall under 9,500 g. Assess the strength of the evidence in the data that the mean generic towel strength is in fact below the 9,500 g target. (Show the whole five-step significance-testing format.)
   (f) Now put yourself in the place of a quality control inspector concerned that the breaking strength be reasonably consistent—i.e., that σ be small. Suppose in fact it is desirable that σ be no more than 400 g. Use the significance-testing format and assess the strength of the evidence given in the data that in fact σ exceeds the target standard deviation.
2. Consider the situation of Example 1 in Chapter 1.
   (a) Use the five-step significance-testing format to assess the strength of the evidence collected in this study to the effect that the laying method is superior to the hanging method in terms of mean runouts produced.
   (b) Make and interpret 90% two-sided and one-sided confidence intervals for the improvement in mean runout produced by the laying method over the hanging method (for the one-sided interval, give a lower bound for $\mu_{\text{hung}} - \mu_{\text{laid}}$).
   (c) Make and interpret a 90% two-sided confidence interval for the mean runout for laid gears.

(d) What is it about Figure 1.1 that makes it questionable whether "normal distribution" prediction and tolerance interval formulas ought to be used to describe runouts for laid gears? Suppose instead that you used the methods of Section 6.6.3 to make prediction and tolerance intervals for laid gear runouts. What confidence could be associated with the largest observed laid runout as an upper prediction bound for a single additional laid runout? What confidence could be associated with the largest observed laid runout as an upper tolerance bound for 95% of additional laid gear runouts?

3. Consider the situation of Example 1 in Chapter 4. In particular, limit attention to those densities obtained under the 2,000 and 4,000 psi pressures. (One can view the six corresponding densities as two samples of size $n_1 = n_2 = 3$.)

(a) Assess the strength of the evidence that increasing pressure increases the mean density of the resulting cylinders. Use the five-step significance-testing format.

(b) Give a 99% lower confidence bound for the increase in mean density associated with the change from 2,000 to 4,000 psi conditions.

(c) Assess the strength of the evidence (in the six density values) that the variability in density differs for the 2,000 and 4,000 psi conditions (i.e., that $\sigma_{2,000} \neq \sigma_{4,000}$).

(d) Give a 90% two-sided confidence interval for the ratio of density standard deviations for the two pressures.

(e) What model assumptions stand behind the formal inferences you made in parts (a) through (d) above?

4. Simple counting with the data of Chapter Exercise 2 in Chapter 3 shows that 18 out of the 26 PET samples had aluminum contents above 100 ppm. Give a two-sided approximate 95% confidence interval for the fraction of all such samples with aluminum contents above 100 ppm.

5. Losen, Cahoy, and Lewis measured the lengths of some spanner bushings of a particular type purchased from a local machine supply shop. The lengths obtained by one of the students were as follows (the units are inches):

1.1375, 1.1390, 1.1420, 1.1430, 1.1410, 1.1360, 1.1395, 1.1380, 1.1350, 1.1370, 1.1345, 1.1340, 1.1405, 1.1340, 1.1380, 1.1355

(a) If you were to, for example, make a confidence interval for the population mean measured length of these bushings via the formulas in Section 6.3, what model assumption must you employ? Make a probability plot to assess the reasonableness of the assumption.

(b) Make a 90% two-sided confidence interval for the mean measured length for bushings of this type measured by this student.

(c) Give an upper bound for the mean length with 90% associated confidence.

(d) Make a 90% two-sided prediction interval for a single additional measured bushing length.

(e) Make a 95% two-sided tolerance interval for 99% of additional measured bushing lengths.

(f) Consider the statistical interval derived from the minimum and maximum sample values—namely, (1.1340, 1.1430). What confidence level should be associated with this interval as a prediction interval for a single additional bushing length? What confidence level should be associated with this interval as a tolerance interval for 99% of additional bushing lengths?

6. The study mentioned in Exercise 5 also included measurement of the outside diameters of the 16 bushings. Two of the students measured each of the bushings, with the results given here.

Bushing	**1**	**2**	**3**	**4**
Student A	.3690	.3690	.3690	.3700
Student B	.3690	.3695	.3695	.3695
Bushing	**5**	**6**	**7**	**8**
Student A	.3695	.3700	.3695	.3690
Student B	.3695	.3700	.3700	.3690

Bushing	9	10	11	12
Student A	.3690	.3695	.3690	.3690
Student B	.3700	.3690	.3695	.3695
Bushing	**13**	**14**	**15**	**16**
Student A	.3695	.3700	.3690	.3690
Student B	.3690	.3695	.3690	.3690

(a) If you want to compare the two students' average measurements, the methods of formulas (6.35), (6.36), and (6.38) are not appropriate. Why?
(b) Make a 95% two-sided confidence interval for the mean difference in outside diameter measurements for the two students.

7. Find the following quantiles using the tables of Appendix B:
(a) the .90 quantile of the t_5 distribution
(b) the .10 quantile of the t_5 distribution
(c) the .95 quantile of the χ^2_7 distribution
(d) the .05 quantile of the χ^2_7 distribution
(e) the .95 quantile of the F distribution with numerator degrees of freedom 8 and denominator degrees of freedom 4
(f) the .05 quantile of the F distribution with numerator degrees of freedom 8 and denominator degrees of freedom 4

8. Find the following quantiles using the tables of Appendix B:
(a) the .99 quantile of the t_{13} distribution
(b) the .01 quantile of the t_{13} distribution
(c) the .975 quantile of the χ^2_3 distribution
(d) the .025 quantile of the χ^2_3 distribution
(e) the .75 quantile of the F distribution with numerator degrees of freedom 6 and denominator degrees of freedom 12
(f) the .25 quantile of the F distribution with numerator degrees of freedom 6 and denominator degrees of freedom 12

9. Ho, Lewer, Peterson, and Riegel worked with the lack of flatness in a particular kind of manufactured steel disk. Fifty different parts of this type were measured for what the students called "wobble," with the results that the 50 (positive) values obtained had mean $\bar{x} = .0287$ in. and standard deviation $s = .0119$ in.
(a) Give a 95% two-sided confidence interval for the mean wobble of all such disks.
(b) Give a lower bound for the mean wobble possessing a 95% confidence level.
(c) Suppose that these disks are ordered with the requirement that the mean wobble not exceed .025 in. Assess the strength of the evidence in the students' data to the effect that the requirement is being violated. Show the whole five-step format.
(d) Is the requirement of part (c) the same as an upper specification of .025 in. on individual wobbles? Explain. Is it possible for a lot with many individual wobbles exceeding .025 in. to meet the requirement of part (c)?
(e) Of the measured wobbles, 19 were .030 in. or more. Use this fact and make an approximate 90% two-sided confidence interval for the fraction of all such disks with wobbles of at least .030 in.

10. T. Johnson tested properties of several brands of 10 lb test monofilament fishing line. Part of his study involved measuring the stretch of a fixed length of line under a 3.5 kg load. Test results for three pieces of two of the brands follow. The units are cm.

Brand B	Brand D
.86, .88, .88	1.06, 1.02, 1.04

(a) Considering first only Brand B, use "normal distribution" model assumptions and give a 90% upper prediction bound for the stretch of an additional piece of Brand B line.
(b) Again considering only Brand B, use "normal distribution" model assumptions and give a 95% upper tolerance bound for stretch measurements of 90% of such pieces of Brand B line.
(c) Again considering only Brand B, use "normal distribution" model assumptions and give 90% two-sided confidence intervals for the

mean and for the standard deviation of the Brand B stretch distribution.

(d) Compare the Brand B and Brand D standard deviations of stretch using an appropriate 90% two-sided confidence interval.

(e) Compare the Brand B and Brand D mean stretch values using an appropriate 90% two-sided confidence interval. Does this interval give clear indication of a difference in mean stretch values for the two brands?

(f) Carry out a formal significance test of the hypothesis that the two brands have the same mean stretch values (use a two-sided alternative hypothesis). Does the conclusion you reach here agree with your answer to part (e)?

11. The accompanying data are $n = 10$ daily measurements of the purity (in percent) of oxygen being delivered by a certain industrial air products supplier. (These data are similar to some given in a November 1990 article in *Chemical Engineering Progress* and used in Chapter Exercise 10 of Chapter 3.)

99.77	99.66	99.61	99.59	99.55
99.64	99.53	99.68	99.49	99.58

(a) Make a normal plot of these data. What does the normal plot reveal about the shape of the purity distribution? ("It is not bell-shaped" is not an adequate answer. Say how its shape departs from the normal shape.)

(b) What statistical "problems" are caused by lack of a normal distribution shape for data such as these?

As a way to deal with problems like those from part (b), you might try transforming the original data. Next are values of $y' = \ln(y - 99.3)$ corresponding to each of the original data values y, and some summary statistics for the transformed values.

$-.76$	-1.02	-1.17	-1.24	-1.39
-1.08	-1.47	$-.97$	-1.66	-1.27

$$\bar{y}' = -1.203 \quad \text{and} \quad s_{y'} = .263$$

(c) Make a normal plot of the transformed values and verify that it is very linear.

(d) Make a 95% two-sided prediction interval for the next transformed purity delivered by this supplier. What does this "untransform" to in terms of raw purity?

(e) Make a 99% two-sided tolerance interval for 95% of additional transformed purities from this supplier. What does this "untransform" to in terms of raw purity?

(f) Suppose that the air products supplier advertises a median purity of at least 99.5%. This corresponds to a median (and therefore mean) transformed value of at least -1.61. Test the supplier's claim ($H_0: \mu_{y'} = -1.61$) against the possibility that the purity is substandard. Show and carefully label all five steps.

12. Chapter Exercise 6 of Chapter 3 contains a data set on the lifetimes (in numbers of 24 mm deep holes drilled in 1045 steel before tool failure) of 12 D952-II (8 mm) drills. The data there have mean $\bar{y} = 117.75$ and $s = 51.1$ holes drilled. Suppose that a normal distribution can be used to roughly describe drill lifetimes.

(a) Give a 90% lower confidence bound for the mean lifetime of drills of this type in this kind of industrial application.

(b) Based on your answer to (a), do you think a hypothesis test of $H_0: \mu = 100$ versus $H_a: \mu > 100$ would have a large p-value or a small p-value? Explain.

(c) Give a 90% lower prediction bound for the next life length of a drill of this type in this kind of industrial application.

(d) Give two-sided tolerance limits with 95% confidence for 90% of all life lengths for drills of this type in this kind of industrial application.

(e) Give two-sided 90% confidence limits for the standard deviation of life lengths for drills of this type in this kind of industrial application.

13. M. Murphy recorded the mileages he obtained while commuting to school in his nine-year-old economy car. He kept track of the mileage for ten

different tankfuls of fuel, involving gasoline of two different octanes. His data follow.

87 Octane	90 Octane
26.43, 27.61, 28.71, 28.94, 29.30	30.57, 30.91, 31.21, 31.77, 32.86

(a) Make normal plots for these two samples of size 5 on the same set of axes. Does the "equal variances, normal distributions" model appear reasonable for describing this situation?
(b) Find s_P for these data. What is this quantity measuring in the present context?
(c) Give a 95% two-sided confidence interval for the difference in mean mileages obtainable under these circumstances using the fuels of the two different octanes. From the nature of this confidence interval, would you expect to find a large p-value or a small p-value when testing H_0: $\mu_{87} = \mu_{90}$ versus H_a: $\mu_{87} \neq \mu_{90}$?
(d) Conduct a significance test of H_0: $\mu_{87} = \mu_{90}$ against the alternative that the higher-octane gasoline provides a higher mean mileage.
(e) Give 95% lower prediction bounds for the next mileages experienced, using first 87 octane fuel and then 90 octane fuel.
(f) Give 95% lower tolerance bounds for 95% of additional mileages experienced, using first 87 octane fuel and then 90 octane fuel.

14. Eastman, Frye, and Schnepf worked with a company that mass-produces plastic bags. They focused on start-up problems of a particular machine that could be operated at either a high speed or a low speed. One part of the data they collected consisted of counts of faulty bags produced in the first 250 manufactured after changing a roll of plastic feedstock. The counts they obtained for both low- and high-speed operation of the machine were 147 faulty ($\hat{p}_H = \frac{147}{250}$) under high-speed operation and 12 faulty under low-speed operation ($\hat{p}_L = \frac{12}{250}$). Suppose that it is sensible to think of the machine as operating in a physically stable fashion during the production of the first 250 bags after changing a roll of plastic, with a constant probability (p_H or p_L) of any particular bag produced being faulty.
(a) Give a 95% upper confidence bound for p_H.
(b) Give a 95% upper confidence bound for p_L.
(c) Compare p_H and p_L using an appropriate two-sided 95% confidence interval. Does this interval provide a clear indication of a difference in the effectiveness of the machine at start-up when run at the two speeds? What kind of a p-value (big or small) would you expect to find in a test of H_0: $p_H = p_L$ versus H_a: $p_H \neq p_L$?
(d) Use the five-step format and test H_0: $p_H = p_L$ versus H_a: $p_H \neq p_L$.

15. Hamilton, Seavey, and Stucker measured resistances, diameters, and lengths for seven copper wires at two different temperatures and used these to compute experimental resistivities for copper at these two temperatures. Their data follow. The units are 10^{-8} Ωm.

Wire	0.0°C	21.8°C
1	1.52	1.72
2	1.44	1.56
3	1.52	1.68
4	1.52	1.64
5	1.56	1.69
6	1.49	1.71
7	1.56	1.72

(a) Suppose that primary interest here centers on the difference between resistivities at the two different temperatures. Make a normal plot of the seven observed differences. Does it appear that a normal distribution description of the observed difference in resistivities at these two temperatures is plausible?
(b) Give a 90% two-sided confidence interval for the mean difference in resistivity measurements for copper wire of this type at 21.8°C and 0.0°C.

(c) Give a 90% two-sided prediction interval for an additional difference in resistivity measurements for copper wire of this type at 21.8°C and 0.0°C.

16. The students referred to in Exercise 15 also measured the resistivities for seven aluminum wires at the same temperatures. The 21.8°C measurements that they obtained follow:

2.65, 2.83, 2.69, 2.73, 2.53, 2.65, 2.69

(a) Give a 99% two-sided confidence interval for the mean resistivity value derived from such experimental determinations.
(b) Give a 95% two-sided prediction interval for the next resistivity value that would be derived from such an experimental determination.
(c) Give a 95% two-sided tolerance interval for 99% of resistivity values derived from such experimental determinations.
(d) Give a 95% two-sided confidence interval for the standard deviation of resistivity values derived from such experimental determinations.
(e) How strong is the evidence that there is a real difference in the precisions with which the aluminum resistivities and the copper resistivities can be measured at 21.8°C? (Carry out a significance test of $H_0: \sigma_{\text{copper}} = \sigma_{\text{aluminum}}$ versus $H_a: \sigma_{\text{copper}} \neq \sigma_{\text{aluminum}}$ using the data of this problem and the 21.8°C data of Exercise 15.)
(f) Again using the data of this exercise and Exercise 15, give a 90% two-sided confidence interval for the ratio $\sigma_{\text{copper}}/\sigma_{\text{aluminum}}$.

17. **(The Stein Two-Stage Estimation Procedure)** One of the most common of all questions faced by engineers planning a data-based study is how much data to collect. The last part of Example 3 illustrates a rather crude method of producing an answer to the sample-size question when estimation of a single mean is involved. In fact, in such circumstances, a more careful two-stage procedure due to Charles Stein can sometimes be used to find appropriate sample sizes.

Suppose that one wishes to use an interval of the form $\bar{x} \pm \Delta$ with a particular confidence coefficient to estimate the mean μ of a normal distribution. If it is desirable to have $\Delta \leq \#$ for some number # and one can collect data in two stages, it is possible to choose an overall sample size to satisfy these criteria as follows. After taking a small or moderate initial sample of size n_1 (n_1 must be at least 2 and is typically at least 4 or 5), one computes the sample standard deviation of the initial data—say, s_1. Then if t is the appropriate t_{n_1-1} distribution quantile for producing the desired (one- or two-sided) confidence, it is necessary to find the smallest integer n such that

$$n \geq \left(\frac{ts_1}{\#}\right)^2$$

If this integer is larger than n_1, then $n_2 = n - n_1$ additional observations are taken. (Otherwise, $n_2 = 0$.) Finally, with $\bar{x}$ the sample mean of all the observations (from both the initial and any subsequent sample), the formula $\bar{x} \pm ts_1/\sqrt{n_1 + n_2}$ (with t still based on $n_1 - 1$ degrees of freedom) is used to estimate μ.

Suppose that in estimating the mean resistance of a production run of resistors, it is desirable to have the two-sided confidence level be 95% and the "± part" of the interval no longer than .5 Ω.
(a) If an initial sample of $n_1 = 5$ resistors produces a sample standard deviation of 1.27 Ω, how many (if any) additional resistors should be sampled in order to meet the stated goals?
(b) If all of the $n_1 + n_2$ resistors taken together produce the sample mean $\bar{x} = 102.8$ Ω, what confidence interval for μ should be declared?

18. Example 15 of Chapter 5 concerns some data on service times at a residence hall depot counter. The data portrayed in Figure 5.21 are decidedly nonnormal-looking, so prediction and tolerance interval formulas based on normal distributions are not appropriate for use with these data. However, the largest of the $n = 65$ observed service times in that figure is 87 sec.

(a) What prediction confidence level can be associated with 87 sec as an upper prediction bound for a single additional service time?
(b) What confidence level can be associated with 87 sec as an upper tolerance bound for 95% of service times?

19. Caliste, Duffie, and Rodriguez studied the process of keymaking using a manual machine at a local lumber yard. The records of two different employees who made keys during the study period were as follows. Employee 1 made a total of 54 different keys, 5 of which were returned as not fitting their locks. Employee 2 made a total of 73 different keys, 22 of which were returned as not fitting their locks.
(a) Give approximate 95% two-sided confidence intervals for the long-run fractions of faulty keys produced by these two different employees.
(b) Give an approximate 95% two-sided confidence interval for the difference in long-run fractions of faulty keys produced by these two different employees.
(c) Assess the strength of the evidence provided in these two samples of a real difference in the keymaking proficiencies of these two employees. (Test H_0: $p_1 = p_2$ using a two-sided alternative hypothesis.)

20. The article "Optimizing Heat Treatment with Factorial Design" by T. Lim (*JOM*, 1989) discusses the improvement of a heat-treating process for gears through the use of factorial experimentation. To compare the performance of the heat-treating process under the original settings of process variables to that using the "improved" settings (identified through factorial experimentation), $n_1 = n_2 = 10$ gears were treated under both sets of conditions. Then measures of flatness, y_1 (in mm of distortion), and concentricity, y_2 (again in mm of distortion), were made on each of the gears. The data shown were read from graphs in the article (and may in some cases differ by perhaps $\pm .002$ mm from the original measurements).

Improved settings

Gear	y_1 (mm)	y_2 (mm)
1A	.036	.050
2A	.040	.054
3A	.026	.043
4A	.051	.071
5A	.034	.043
6A	.050	.058
7A	.059	.061
8A	.055	.048
9A	.051	.060
10A	.050	.033

Original settings

Gear	y_1 (mm)	y_2 (mm)
1B	.056	.070
2B	.064	.062
3B	.070	.075
4B	.037	.060
5B	.054	.071
6B	.060	.070
7B	.065	.060
8B	.060	.060
9B	.051	.070
10B	.062	.070

(a) What assumptions are necessary in order to make inferences regarding the parameters of the y_1 (or y_2) distribution for the improved settings of the process variables?
(b) Make a normal plot for the improved settings' y_1 values. Does it appear that it is reasonable to treat the improved settings' flatness distribution as normal? Explain.
(c) Suppose that the improved settings' flatness distribution is normal, and do the following:
(i) Give a 90% two-sided confidence interval for the mean flatness distortion value for gears of this type.
(ii) Give a 90% two-sided prediction interval for an additional flatness distortion value.

(iii) Give a 95% two-sided tolerance interval for 90% of additional flatness distortion values.

(iv) Give a 90% two-sided confidence interval for the standard deviation of flatness distortion values for gears of this type.

(d) Repeat parts (b) and (c) using the improved settings' concentricity values, y_2, instead of flatness.

(e) Explain why it is not possible to base formal inferences (tests and confidence intervals), for comparing the standard deviations of the y_1 and y_2 distributions for the improved process settings, on the sample standard deviations of the y_1 and y_2 measurements from gears 1A through 10A.

(f) What assumptions are necessary in order to make comparisons between parameters of the y_1 (or y_2) distributions for the original and improved settings of the process variables?

(g) Make normal plots of the y_1 data for the original settings and for the improved settings on the same set of axes. Does an "equal variances, normal distributions" model appear tenable here? Explain.

(h) Supposing that the flatness distortion distributions for the original and improved process settings are adequately described as normal with a common standard deviation, do the following.

(i) Use an appropriate significance test to assess the strength of the evidence in the data to the effect that the improved settings produce a reduction in mean flatness distortion.

(ii) Give a 90% lower confidence bound on the reduction in mean flatness distortion provided by the improved process settings.

(i) Repeat parts (g) and (h) using the y_2 values and concentricity instead of flatness.

21. R. Behne measured air pressure in car tires in a student parking lot. Shown here is one summary of the data he reported. Any tire with pressure reading more than 3 psi below its recommended value was considered underinflated, while any tire with pressure reading more than 3 psi above its recommended value was considered overinflated. The counts in the accompanying table are the numbers of cars (out of 25 checked) falling into the four possible categories.

		Underinflated tires	
		None	At Least One Tire
Overinflated tires	None	6	5
	At Least One Tire	10	4

(a) Behne's sample was in all likelihood a convenience sample (as opposed to a genuinely simple random sample) of the cars in the large lot. Does it make sense to argue in this case that the data can be treated as if the sample were a simple random sample? On what basis? Explain.

(b) Give a two-sided 90% confidence interval for the fraction of all cars in the lot with at least one underinflated tire.

(c) Give a two-sided 90% confidence interval for the fraction of the cars in the lot with at least one overinflated tire.

(d) Give a 90% lower confidence bound on the fraction of cars in the lot with at least one misinflated tire.

(e) Why can't the data here be used with formula (6.67) of Section 6.5 to make a confidence interval for the difference in the fraction of cars with at least one underinflated tire and the fraction with at least one overinflated tire?

22. The article "A Recursive Partitioning Method for the Selection of Quality Assurance Tests" by Raz and Bousum (*Quality Engineering*, 1990) contains some data on the fractions of torque converters manufactured in a particular facility failing a final inspection (and thus requiring some rework). For a particular family of four-element converters, about 39% of 442 converters tested were out of specifications on a high-speed operation inlet flow test.

(a) If plant conditions tomorrow are like those under which the 442 converters were manufactured, give a two-sided 98% confidence interval for the probability that a given converter manufactured will fail the high-speed inlet flow test.

(b) Suppose that a process change is instituted in an effort to reduce the fraction of converters failing the high-speed inlet flow test. If only 32 out of the first 100 converters manufactured fail the high-speed inlet flow test, is this convincing evidence that a real process improvement has been accomplished? (Give and interpret a 90% two-sided confidence interval for the change in test failure probability.)

23. Return to the situation of Chapter Exercise 1 in Chapter 3 and the measured gains of 120 amplifiers. The nominal/design value of the gain was 10.0 dB; 16 of the 120 amplifiers measured had gains above nominal. Give a 95% two-sided confidence interval for the fraction of all such amplifiers with above-nominal gains.

24. The article "Multi-functional Pneumatic Gripper Operating Under Constant Input Actuation Air Pressure" by J. Przybyl (*Journal of Engineering Technology*, 1988) discusses the performance of a 6-digit pneumatic robotic gripper. One part of the article concerns the gripping pressure (measured by manometers) delivered to objects of different shapes for fixed input air pressures. The data given here are the measurements (in psi) reported for an actuation pressure of 40 psi for (respectively) a 1.7 in. $\times$ 1.5 in. $\times$ 3.5 in. rectangular bar and a circular bar of radius 1.0 in. and length 3.5 in.

Rectangular Bar	Circular Bar
76	84
82	87
85	94
88	80
82	92

(a) Compare the variabilities of the gripping pressures delivered to the two different objects using an appropriate 98% two-sided confidence interval. Does there appear to be much evidence in the data of a difference between these? Explain.

(b) Supposing that the variabilities of gripping pressure delivered by the gripper to the two different objects are comparable, give a 95% two-sided confidence interval for the difference in mean gripping pressures delivered.

(c) The data here came from the operation of a single prototype gripper. Why would you expect to see more variation in measured gripping pressures than that represented here if each measurement in a sample were made on a different gripper? Strictly speaking, to what do the inferences in (a) and (b) apply? To the single prototype gripper or to all grippers of this design? Discuss this issue.

25. A sample of 95 U-bolts produced by a small company has thread lengths with a mean of $\bar{x} = 10.1$ (.001 in. above nominal) and $s = 3.2$ (.001 in.).

(a) Give a 95% two-sided confidence interval for the mean thread length (measured in .001 in. above nominal). Judging from this interval, would you expect a small or a large p-value when testing $H_0: \mu = 0$ versus $H_a: \mu \neq 0$? Explain.

(b) Use the five-step format of Section 6.2 and assess the strength of the evidence provided by the data to the effect that the population mean thread length exceeds nominal.

26. D. Kim did some crude tensile strength testing on pieces of some nominally .012 in. diameter wire of various lengths. Below are Kim's measured strengths (kg) for pieces of wire of lengths 25 cm and 30 cm.

25 cm Lengths	30 cm Lengths
4.00, 4.65, 4.70, 4.50	4.10, 4.50, 3.80, 4.60
4.40, 4.50, 4.50, 4.20	4.20, 4.60, 4.60, 3.90

(a) If one is to make a confidence interval for the mean measured strength of 25 cm pieces of this wire using the methods of Section 6.3, what model assumption must be employed? Make a probability plot useful in assessing the reasonableness of the assumption.
(b) Make a 95% two-sided confidence interval for the mean measured strength of 25 cm pieces of this wire.
(c) Give a 95% lower confidence bound for the mean measured strength of 25 cm pieces.
(d) Make a 95% two-sided prediction interval for a single additional measured strength for a 25 cm piece of wire.
(e) Make a 99% two-sided tolerance interval for 95% of additional measured strengths of 25 cm pieces of this wire.
(f) Consider the statistical interval derived from the minimum and maximum sample values for the 25 cm lengths—namely, (4.00, 4.70). What confidence should be associated with this interval as a prediction interval for a single additional measured strength? What confidence should be associated with this interval as a tolerance interval for 95% of additional measured strengths for 25 cm pieces of this wire?
(g) In order to make formal inferences about $\mu_{25} - \mu_{30}$ based on these data, what must you be willing to use for model assumptions? Make a plot useful for investigating the reasonableness of those assumptions.
(h) Proceed under the assumptions discussed in part (g) and assess the strength of the evidence provided by Kim's data to the effect that an increase in specimen length produces a decrease in measured strength.
(i) Proceed under the necessary model assumptions to give a 98% two-sided confidence interval for $\mu_{25} - \mu_{30}$.

27. The article "Influence of Final Recrystallization Heat Treatment on Zircaloy-4 Strip Corrosion" by Foster, Dougherty, Burke, Bates, and Worcester (*Journal of Nuclear Materials*, 1990) reported some summary statistics from the measurement of the diameters of 821 particles observed in a bright field TEM micrograph of a Zircaloy-4 specimen. The sample mean diameter was $\bar{x} = .055\ \mu$m, and the sample standard deviation of the diameters was $s = .028\ \mu$m.
(a) The engineering researchers wished to establish from their observation of this single specimen the impact of a certain combination of specimen lot and heat-treating regimen on particle size. Briefly discuss why data such as the ones summarized have serious limitations for this purpose. (*Hints:* The apparent "sample size" here is huge. But of what is there a sample? How widely do the researchers want their results to apply? Given this desire, is the "real" sample size really so large?)
(b) Use the sample information and give a 98% two-sided confidence interval for the mean diameter of particles in this particular Zircaloy-4 specimen.
(c) Suppose that a standard method of heat treating for such specimens is believed to produce a mean particle diameter of .057 μm. Assess the strength of the evidence contained in the sample of diameter measurements to the effect that the specimen's mean particle diameter is different from the standard. Show the whole five-step format.
(d) Discuss, in the context of part (c), the potential difference between the mean diameter being statistically different from .057 μm and there being a difference between μ and .057 that is of practical importance.

28. Return to Kim's tensile strength data given in Exercise 26.
(a) Operating under the assumption that measured tensile strengths of 25 cm lengths of the wire studied are normally distributed, give a two-sided 98% confidence interval for the standard deviation of measured strengths.
(b) Operating under the assumption that measured tensile strengths of 30 cm lengths of the wire studied are normally distributed, give a 95% upper confidence bound for the standard deviation of measured strengths.

(c) Operating under the assumption that both 25 and 30 cm lengths of the wire have normally distributed measured tensile strengths, assess the strength of Kim's evidence that 25 and 30 cm lengths differ in variability of their measured tensile strengths. (Use $H_0: \sigma_{25} = \sigma_{30}$ and $H_a: \sigma_{25} \neq \sigma_{30}$ and show the whole five-step format.)

(d) Operating under the assumption that both 25 and 30 cm lengths produce normally distributed tensile strengths, give a 98% two-sided confidence interval for the ratio σ_{25}/σ_{30}.

29. Find the following quantiles:

(a) the .99 quantile of the χ^2_4 distribution

(b) the .025 quantile of the χ^2_4 distribution

(c) the .99 quantile of the F distribution with numerator degrees of freedom 3 and denominator degrees of freedom 15

(d) the .25 quantile of the F distribution with numerator degrees of freedom 3 and denominator degrees of freedom 15

30. The digital and vernier caliper measurements of no. 10 machine screw diameters summarized in Exercise 3 of Section 6.3 are such that for 19 out of 50 of the screws, there was no difference in the measurements. Based on these results, give a 95% confidence interval for the long-run fraction of such measurements by the student technician that would produce agreement between the digital and vernier caliper measurements.

31. Duren, Leng, and Patterson studied the drilling of holes in a miniature metal part using electrical discharge machining. Blueprint specifications on a certain hole called for diameters of $.0210 \pm .0003$ in. The diameters of this hole were measured on 50 parts with plug gauges and produced $\bar{x} = .02046$ and $s = .00178$. Assume that the holes the students measured were representative of the output of a physically stable drilling process.

(a) Give a 95% two-sided confidence interval for the mean diameter of holes drilled by this process.

(b) Give a 95% lower confidence bound for the mean diameter of the holes drilled by this process. (Find a number, #, so that $(\#, \infty)$ is a 95% confidence interval.) How does this number compare to the lower end point of your interval from (a)?

(c) Repeat (a) using 90% confidence. How does this interval compare with the one from (a)?

(d) Repeat (b) using 90% confidence. How does this bound compare to the one found in (b)?

(e) Interpret your interval from (a) for someone with little statistical background. (Speak in the context of the drilling study and use the "authorized interpretation" of confidence as your guide.)

(f) Based on your confidence intervals, would you expect the p-value in a test of $H_0: \mu = .0210$ versus $H_a: \mu \neq .0210$ to be small? Explain.

(g) Based on your confidence intervals, would you expect the p-value in a test of $H_0: \mu = .0210$ versus $H_a: \mu > .0210$ to be small? Explain.

(h) Consider again your answer to part (a). A colleague sees your calculations and says, "Oh, so 95% of the measured diameters would be in that range?" What do you say to this person?

(i) Use the five step significance-testing format of Section 6.2 and assess the strength of the evidence provided by the data to the effect that the process mean diameter differs from the mid-specification of .0210. (Begin with $H_0: \mu = .0210$ and use $H_a: \mu \neq .0210$.

(j) Thus far in this exercise, inference for the mean hole diameter has been of interest. Explain why in practice the variability of diameters is also important. The methods of Sections 6.1 are not designed for analyzing distributional spread. Where in Chapter 6 can you find inference methods for this feature?

32. Return to Babcock's fatigue life testing data in Chapter Exercise 18 of Chapter 3 and for now focus on the fatigue life data for heat 1.

(a) In order to do inference based on this small sample, what model assumptions must you employ? What does a normal plot say about the appropriateness of these assumptions?

(b) Give a 90% two-sided confidence interval for the mean fatigue life of such specimens from this heat.
(c) Give a 90% lower confidence bound for the mean fatigue life of such specimens from this heat.
(d) If you are interested in quantifying the variability in fatigue lives produced by this heat of steel, inference for σ becomes relevant. Give a 95% two-sided confidence interval for σ based on display (6.42) of the text.
(e) Make a 90% two-sided prediction interval for a single additional fatigue life for a specimen from this heat.
(f) Make a 95% two-sided tolerance interval for 90% of additional fatigue lives for specimens from this heat. How does this interval compare to your interval from (e)?
(g) Now consider the statistical interval derived from the minimum and maximum sample values from heat 1, namely (11, 548). What confidence should be associated with this interval as a prediction interval for a single additional fatigue life from this heat? What confidence should be associated with the interval as a tolerance interval for 90% of additional fatigue lives?

Now consider both the data for heat 1 and the data for heat 3.

(h) In order to make formal inferences about $\mu_1 - \mu_3$ based on these data, what must be assumed about fatigue lives for specimens from these two heats? Make a plot useful for investigating the reasonableness of these assumptions.
(i) Under the appropriate assumptions (state them), give a 95% two-sided confidence interval for $\mu_1 - \mu_3$.

33. Consider the Notch/Dial Bore and Notch/Air Spindler measurements on ten servo sleeves recorded in Chapter Exercise 19 in Chapter 3.

(a) If one wishes to compare the dial bore gauge and the air spindler gauge measurements, the methods of formulas (6.35), (6.36), and (6.38) are not appropriate. Why?
(b) What assumption must you make in order to do formal inference on the mean difference in dial bore and air spindler gauge measurements? Make a plot useful for assessing the reasonableness of this assumption. Comment on what it indicates in this problem.
(c) Make the necessary assumptions about the dial bore and air spindler measurements and assess the strength of the evidence in the data of a systematic difference between the two gauges.
(d) Make a 95% two-sided confidence interval for the mean difference in dial bore and air spindler measurements.
(e) Briefly discuss how your answers for parts (c) and (d) of this problem are consistent.

34. Chapter Exercise 20 in Chapter 3 concerned the drilling of holes in miniature metal parts using laser drilling and electrical discharge machining. Return to that problem and consider first only the EDM values.

(a) In order to use the methods of inference of Section 6.3 with these data, what model assumptions must be made? Make a plot useful for investigating the appropriateness of those assumptions. Comment on the shape of that plot and what it says about the appropriateness of the model assumptions.
(b) Give a 99% two-sided confidence interval for the mean angle produced by the EDM drilling of this hole.
(c) Give a 99% upper confidence bound for the mean angle produced by the EDM drilling of this hole.
(d) Give a 95% two-sided confidence interval for the standard deviation of angles produced by the EDM drilling of this hole.
(e) Make a 99% two-sided prediction interval for the next measured angle produced by the EDM drilling of this hole.
(f) Make a 95% two-sided tolerance interval for 99% of angles produced by the EDM drilling of this hole.
(g) Consider the statistical interval derived from the minimum and maximum sample EDM

values, namely (43.2, 46.1). What confidence should be associated with this interval as a prediction interval for a single additional measured angle? What confidence should be associated with this interval as a tolerance interval for 99% of additional measured angles?

Now consider both the EDM and initial set of Laser values in Chapter Exercise 20 of Chapter 3 (two sets of 13 parts).

(h) In order to make formal inferences about $\mu_{\text{Laser}} - \mu_{\text{EDM}}$ based on these data, what must you be willing to use for model assumptions? Make a plot useful for investigating the reasonableness of those assumptions.

(i) Proceed under appropriate assumptions to assess the strength of the evidence provided by the data that there is a difference in the mean angles produced by the two drilling methods.

(j) Give a 95% two-sided confidence interval for $\mu_{\text{Laser}} - \mu_{\text{EDM}}$.

(k) Give a 90% two-sided confidence interval for comparing the standard deviations of angles produced by Laser and EDM drilling of this hole.

Now consider both sets of Laser measurements given in Chapter Exercise 20 of Chapter 3. (Holes A and B are on the same 13 parts.)

(l) If you wished to compare the mean angle measurements for the two holes, the formulas used in (i) and (j) are not appropriate. Why?

(m) Make a 90% two-sided confidence interval for the mean difference in angles for the two holes made with the laser equipment.

(n) Assess the strength of the evidence provided by these data that there is a systematic difference in the angles of the holes made with the laser equipment.

(o) Briefly discuss why your answers to parts (m) and (n) of this exercise are compatible. (Discuss how the outcome of part (n) could have been anticipated from the outcome of part (m).)

35. A so-called "tilttable" test was run in order to determine the angles at which certain vehicles experience lift-off of one set of wheels and begin to roll over on their sides. "Tilttable ratios" (which are the tangents of the angles at which lift-off occurred) were measured for two minivans of different makes four times each with the following results.

Van 1	Van 2
1.096, 1.093, 1.090, 1.093	.962, .970, .967, .966

(a) If you were to make a confidence interval for the long-run mean measured tilttable ratio for Van 1 (under conditions like those experienced during the testing) using the methods of Section 6.3, what model assumption must be made?

(b) Make a 95% two-sided confidence interval for the mean measured tilttable ratio for Van 1 under conditions like those experienced during the testing.

(c) Give a 95% lower confidence bound for the mean measured tilttable ratio for Van 1.

(d) Give a 95% lower confidence bound for the standard deviation of tilttable ratios for Van 1.

(e) Make a 95% two-sided prediction interval for a single additional measured tilttable ratio for Van 1 under conditions such as those experienced during testing.

(f) Make a 99% two-sided tolerance interval for 95% of additional measured tilttable ratios for Van 1.

(g) Consider the statistical interval derived from the minimum and maximum sample values for Van 1, namely (1.090, 1.096). What confidence should be associated with this interval as a prediction interval for a single additional measured tilttable ratio? What confidence should be associated with this interval as a tolerance interval for 95% of additional tilttable test results for Van 1?

Now consider the data for both vans.

(h) In order to make formal inferences about $\mu_1 - \mu_2$ based on these data, what must you be willing to use for model assumptions?

(i) Proceed under the necessary assumptions to assess the strength of the evidence provided by the data that there is a difference in mean measured tilttable ratios for the two vans.

(j) Proceed under the necessary model assumptions to give a 90% two-sided confidence interval for $\mu_1 - \mu_2$.

(k) Proceed under the necessary model assumptions to give a 90% two-sided confidence interval for σ_1/σ_2.

Chapter 6 Summary Tables

The methods presented in Chapter 6 can seem overwhelming in their variety. It is sometimes helpful to have a summary of them. The tables here give such a summary and can be used to help you locate methods appropriate in a particular problem or application.

Table 1
Inference Methods for Individual Values

Inference For	Assumptions	Interval	Section
x_{n+1} (a single additional value)		$(\min(x_1, \ldots, x_n), \max(x_1, \ldots, x_n))$ or $(\min(x_1, \ldots, x_n), \infty)$ or $(-\infty, \max(x_1, \ldots, x_n))$	6.6
	observations normal	$\bar{x} \pm ts\sqrt{1 + \frac{1}{n}}$	6.6
most of the distribution		$(\min(x_1, \ldots, x_n), \max(x_1, \ldots, x_n))$ or $(\min(x_1, \ldots, x_n), \infty)$ or $(-\infty, \max(x_1, \ldots, x_n))$	6.6
	observations normal	$\bar{x} \pm \tau_2 s$ or $(\bar{x} - \tau_1 s, \infty)$ or $(-\infty, \bar{x} + \tau_1 s)$	6.6

Table 2
Inference Methods for One and Two Means

Inference For	Sample Size	Assumptions	H_0, Test Stat, Reference	Interval	Section
μ (one mean)	large n		$H_0: \mu = \#$ $Z = \dfrac{\bar{x} - \#}{s/\sqrt{n}}$ standard normal	$\bar{x} \pm z\dfrac{s}{\sqrt{n}}$	6.1, 6.2
	small n	observations normal	$H_0: \mu = \#$ $T = \dfrac{\bar{x} - \#}{s/\sqrt{n}}$ t with $\nu = n - 1$	$\bar{x} \pm t\dfrac{s}{\sqrt{n}}$	6.3
$\mu_1 - \mu_2$ (difference in means)	large n_1, n_2	independent samples	$H_0: \mu_1 - \mu_2 = \#$ $Z = \dfrac{\bar{x}_1 - \bar{x}_2 - \#}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ standard normal	$\bar{x}_1 - \bar{x}_2 \pm z\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$	6.3
	small n_1 or n_2	independent normal samples $\sigma_1 = \sigma_2$	$H_0: \mu_1 - \mu_2 = \#$ $T = \dfrac{\bar{x}_1 - \bar{x}_2 - \#}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ t with $\nu = n_1 + n_2 - 2$	$\bar{x}_1 - \bar{x}_2 \pm ts_P\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$	6.3
		possibly $\sigma_1 \neq \sigma_2$		$\bar{x}_1 - \bar{x}_2 \pm \hat{t}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ use random $\hat{\nu}$ given in (6.37)	6.3
μ_d (mean difference)	large n	(paired data)	$H_0: \mu_d = \#$ $Z = \dfrac{\bar{d} - \#}{s_d/\sqrt{n}}$ standard normal	$\bar{d} \pm z\frac{s_d}{\sqrt{n}}$	6.3
	small n	(paired data) normal differences	$H_0: \mu_d = \#$ $T = \dfrac{\bar{d} - \#}{s_d/\sqrt{n}}$ t with $\nu = n - 1$	$\bar{d} \pm t\dfrac{s_d}{\sqrt{n}}$	6.3

Table 3
Inference Methods for Variances

Inference For	Assumptions	H_0, Test Stat, Reference	Interval	Section
σ^2 (one variance)	observations normal	$H_0: \sigma^2 = \#$ $X^2 = \dfrac{(n-1)s^2}{\#}$ χ^2 with $\nu = n-1$	$\dfrac{(n-1)s^2}{U}$ and/or $\dfrac{(n-1)s^2}{L}$	6.4
σ_1^2/σ_2^2 (variance ratio)	observations normal independent samples	$H_0: \dfrac{\sigma_1^2}{\sigma_2^2} = \#$ $F = \dfrac{s_1^2/s_2^2}{\#}$ F with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$	$\dfrac{s_1^2}{U \cdot s_2^2}$ and/or $\dfrac{s_1^2}{L \cdot s_2^2}$	6.4

Table 4
Inference Methods for Proportions

Inference For	Sample Size	Assumptions	H_0, Test Stat, Reference	Interval	Section
p (one proportion)	large n		$H_0: p = \#$ $Z = \dfrac{\hat{p} - \#}{\sqrt{\dfrac{\#(1-\#)}{n}}}$ standard normal	$\hat{p} \pm z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ or $\hat{p} \pm z\dfrac{1}{2\sqrt{n}}$	6.5
$p_1 - p_2$ difference in proportions	large n_1, n_2	independent samples	$H_0: p_1 - p_2 = 0$ $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ use $\hat{p}$ given in (6.71) standard normal	$\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ or $\hat{p}_1 - \hat{p}_2 \pm z \cdot \dfrac{1}{2}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$	6.5