Chapter 4

Classical Inference

Classical inference is used to estimate the statistical characteristics of a large population when only a small representative random sample of the population can be obtained. An understanding of classical inference is essential for gaining an appreciation of its strengths and for how Bayesian inference and Dempster-Shafer evidential theory each ameliorate some of its limitations.

Statistical inference uses a number computed from the sample data to make inferences about an unknown number that describes the larger population. In this regard, a *parameter* is a number describing the population and a *statistic* is a number that can be computed from the sample data without using any unknown parameters. The theory discussed in this chapter is applicable when simple random samples can be gathered. A simple random sample of size n consists of n units from the population chosen in such a way that every set of n units has an equal chance to be the sample actually selected.

More elaborate sampling designs are often appropriate. For example, stratified random samples are used to restrict the random selection by dividing the population into groups of similar units called strata. Separate simple random samples are then selected from each stratum, as when sampling geographically dispersed populations. Block sample designs are another way to create a group of experimental units that are known before an experiment begins to be similar in some way that is expected to affect the response to the experiment. In a block design, the random assignment of units to treatments or some other influence is performed separately within each block. A third method of restricting random selection is to perform the selection in stages. This is often done when national samples of families, households, or individuals are required. For example, a multistage sample design for a population survey may be constructed as follows:

Stage 1: gather a sample from the 3,000 counties in the United States;
Stage 2: select a sample of townships within each of the counties chosen;
Stage 3: select a sample of blocks within each chosen township;
Stage 4: gather a sample of households within each block.

Additional information on creating and analyzing the results from these sample designs may be found in the references at the end of this chapter.[1–7]

4.1 Estimating the statistics of a population

The sample mean $\bar{x}$ is an unbiased estimator of an unknown population mean μ if the samples are random and are representative of the entire population. In this case, the standard deviation of the sample mean is

$$\sigma_x = \sigma / \sqrt{n} \,, \tag{4-1}$$

where σ is the standard deviation of the entire population and n is the sample size. The standard deviation of the sample mean is smaller than the standard deviation of the entire population since the standard deviation of the sample mean is obtained by dividing the standard deviation of the population by the square root of the number of observations in the sample.

Figure 4.1 shows that if the random variables that characterize the population are normally distributed, then there is approximately a 68 percent probability that the sample mean is within ±1 standard deviations of the population mean, approximately a 95 percent probability that the sample mean is within ±2 standard deviations of the population mean, and approximately a 99.7 percent probability that the sample mean is within ±3 standard deviations of the population mean.

As an example of how to apply this information, suppose the mean score of a "standardization group" on an aptitude test is 500 and the standard deviation is 100. The scale is maintained from year to year, but the mean in any year can be different than 500. We want to estimate the mean test score for more than 250,000 students using a sample of test scores from 500 students. The test is given to a random sample of 500 students, who get a mean score of 461. What can we say about the mean score of the entire population of 250,000?

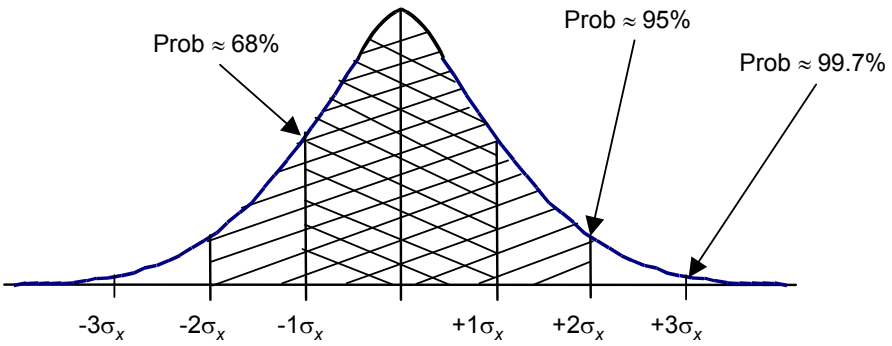

Figure 4.1 Interpretation of the standard deviation of the sample mean for a normal distribution.

The sample maen $\bar{x}$ is 461 and the standard deviation of the sample mean σ_x is $100/\sqrt{500} = 4.5$. Therefore, we can state that we are 95 percent confident that the unknown mean score for the 250,000 students lies between $\bar{x} - 9 = 461 - 9 = 452$ and $\bar{x} + 9 = 461 + 9 = 470$.

The interval $\bar{x} \pm 9$ is the 95 percent *confidence interval* for μ and the *margin of error* is ± 9.

4.2 Interpreting the confidence interval

Confidence intervals have two aspects, the interval computed from the data and the confidence level that gives the probability that the method produces an interval that includes the parameter. Most often a confidence level greater than or equal to 90 percent is selected. If C is the confidence level in decimal form, then a level C confidence interval for a parameter θ is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of θ.

For example, suppose it is desired to find a level C confidence interval for the mean μ of a population from an unbiased random data sample of size n. The confidence interval is based on the sampling distribution for the sample mean $\bar{x}$, which is equal to $N(\mu, \sigma/\sqrt{n})$ when the sample is obtained from a population having the $N(\mu, \sigma)$ distribution. In this notation, N represents a normal distribution, μ the mean of the entire population, and σ the standard deviation of the entire population. The central limit theorem confirms that a normal distribution is a valid representation of the sampling distribution of the sample mean when the sample size is sufficiently large regardless of the probability density function that describes the statistics of the entire population.[7]

The construction of a 95 percent confidence interval is based on the observation that any normal distribution has a probability of 0.95 that the true value of the population mean lies within ± 2 standard deviations of the sample mean. Confidence level C (where C is expressed in decimal form) must include the central area C under the normal curve. To ensure that this area is captured by the confidence level, a number z^* is found such that there is a probability C that a sample from any normal distribution falls within $\pm z^*$ standard deviations of the distribution's mean. The number z^* is listed in tables of standard normal probabilities such as the summary given in Table 4.1.[8]

The value z^* for confidence C encompasses the central area C between $-z^*$ and z^*, thus omitting the area $1 - C$ as illustrated in Figure 4.2. Half the omitted area lies in each tail. Because z^* has area $(1 - C)/2$ to its right under the standard normal curve, it is called the upper $(1 - C)/2$ or p critical value of the standard normal distribution. For example, if $C = 0.95$, there is a $(1 - 0.95)/2$ or 2.5

percent chance that the true population mean is more than two standard deviations larger than the sample mean and an equal probability that it is more than two standard deviations lower than the sample mean. In this case, z^* equal to 1.960 is the upper 2.5 percent critical value for the standard normal distribution.

Table 4.1 Standard normal probabilities showing z^* for various confidence levels.

Confidence Level	$(1 - C)/2$	z^*
90%	0.05	1.645
95%	0.025	1.960
96%	0.02	2.054
98%	0.01	2.326
99%	0.005	2.576
99.5%	0.0025	2.807
99.8%	0.001	3.091
99.9%	0.0005	3.291

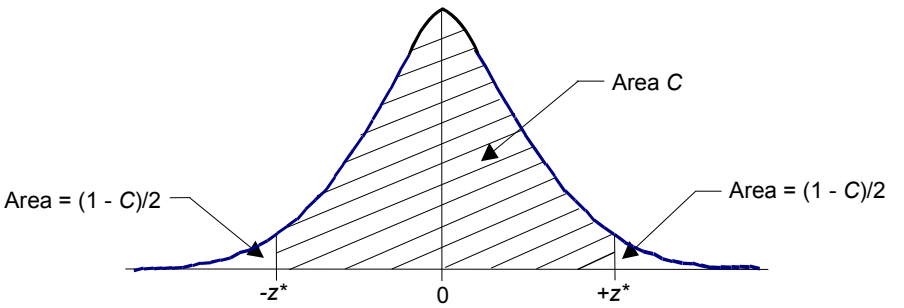

Figure 4.2 Central area of normal distribution included in a confidence level *C*.

Figure 4.3 describes the interpretation of a 95 percent confidence interval in repeated sampling. The center of each interval is marked by a dot. The arrows span the confidence interval. All except 1 of the 25 intervals include the true value of μ. For a large number of samples, 95 percent of the confidence intervals will contain μ.

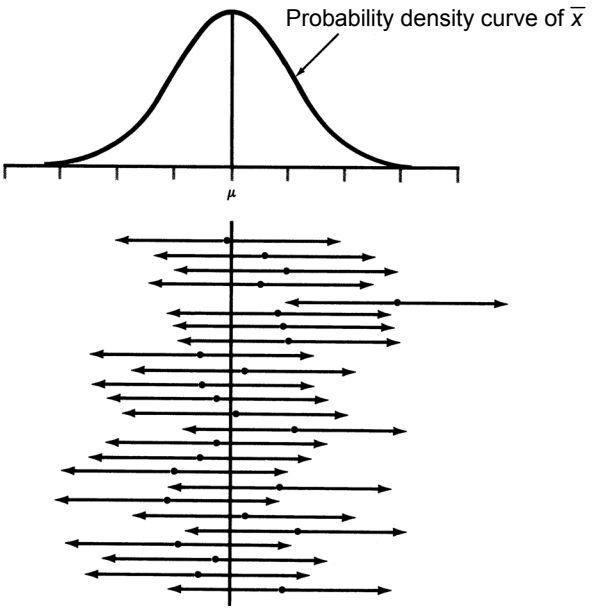

Figure 4.3 Interpretation of confidence interval with repeated sampling. (D.S. Moore and G.P. McCabe, *Introduction to the Practice of Statistics*, 4th Ed., New York, NY: W.H. Freeman and Company [Aug. 2002].)

4.3 Confidence interval for a population mean

If the sample mean $\bar{x}$ is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$, i.e., $N(\mu, \sigma/\sqrt{n})$, the probability is C that $\bar{x}$ lies between

$$\mu - z^*\sigma/\sqrt{n} \text{ and } \mu + z^*\sigma/\sqrt{n} .$$

This is equivalent to stating that the unknown population mean μ lies between

$$\bar{x} - z^*\sigma/\sqrt{n} \text{ and } \bar{x} + z^*\sigma/\sqrt{n} ,$$

or there is a probability C that the interval $\bar{x} \pm z^*\sigma/\sqrt{n}$ contains μ. Therefore, the interval $\bar{x} \pm z^*\sigma/\sqrt{n}$ is the desired confidence interval.

The estimator of the unknown μ is $\bar{x}$ and the margin of error M is

$$M = z^*\sigma/\sqrt{n} . \tag{4-2}$$

Thus, the sample size n needed to obtain a confidence interval with a specified margin of error M is

$$n = (z^*\sigma/M)^2 , \tag{4-3}$$

assuming randomly selected and unbiased samples, a normally distributed unstratified population, and no outliers (i.e., no individual observations that fall well outside the overall pattern of the data).

The requisite sample size increases as the desired level of confidence increases, dispersion of the sample data increases, and the allowable error decreases. The size of the entire population does not influence the sample size as long as the population is much larger than the sample.[8]

The confidence interval is exact when the population distribution is normal and is approximately correct for large n for other distributions by application of the central limit theorem.[8] There is a tradeoff between the confidence level and the margin of error. To obtain higher confidence from the same data requires acceptance of a larger margin of error. Thus, it is more difficult to arrive at the exact value of the mean μ of a highly variable population, which is why the margin of error of a confidence interval increases with σ. The selected confidence interval depends on the application in which the data are used (e.g., aircraft tracking, missile detection, object counting, average vehicle speed measurement, or historical data collection).

The margin of error in a confidence interval indicates the error expected from chance variation in randomized data production. When random samples are not obtained because of omission of some affected groups from the data sampling or non-response from some groups, additional errors are introduced that may be larger than the random sampling error. If the population is not normal and contains extreme outliers or is strongly skewed, the confidence level will be different from C.

The following examples describe how the sample data and confidence interval provide statistical information about the entire population.

Example 1: suppose a laboratory analyzes a specimen three times for the concentration of a particular compound. The analysis procedure has no bias, implying the mean μ of the population of all measurements is the true concentration of the compound in the specimen. The standard deviation of the analysis procedure is known to be 0.0068 g/l.

The three analyses of the specimen yield compound concentrations of 0.8403, 0.8363, and 0.8447 g/l. What are the 90 percent and 99 percent confidence intervals for the true concentration μ?

From the given sample concentration data, the sample mean of the measurements is

$$\bar{x} = (0.8403 + 0.8363 + 0.8447)/3 \text{ g/l} = 0.8404 \text{ g/l}. \quad (4\text{-}4)$$

Table 4.1 shows that for 90 percent confidence, $z^* = 1.645$ and for 99 percent confidence, $z^* = 2.576$.

Therefore, the 90 percent confidence interval for μ is

$$\bar{x} \pm z^* \sigma/\sqrt{n} = 0.8404 \pm 1.645\ (0.0068/\sqrt{3}) \text{ g/l} = 0.8404 \pm 0.0065 \text{ g/l}$$
$$= 0.8339 \text{ g/l}, 0.8469 \text{ g/l}. \quad (4\text{-}5)$$

The 99 percent confidence interval for μ is

$$\bar{x} \pm z^* \sigma/\sqrt{n} = 0.8404 \pm 2.576\ (0.0068/\sqrt{3}) \text{ g/l} = 0.8404 \pm 0.0101 \text{ g/l}$$
$$= 0.8303 \text{ g/l}, 0.8505 \text{ g/l}. \quad (4\text{-}6)$$

Figure 4.4 illustrates the confidence intervals that correspond to the 90 and 99 percent confidence levels. As expected, the 99 percent confidence interval is larger.

Example 2: A confidence interval is required for missile tracking data. Suppose a data point obtained at time interval *t* for the potential update of a missile track is 100 m from the last update made the interval before. Based on historical data for the identified missile type and the tracking system used, it is known that the mean change in missile position between data updates is 90 m. The standard deviation of the position estimate is 3 m. Should the data at time interval *t* be merged with the established track or should a new track be initiated?

If 99 percent confidence that the data at time interval *t* belong to the existing track is desirable, then the confidence interval is given by

$$\mu \pm z^* \sigma/\sqrt{n} = 90 \pm 2.576\ (3/\sqrt{1}) \text{ m} = 82.27 \text{ m}, 97.73 \text{ m} \quad (4\text{-}7)$$

where $z^* = 2.576$.

Thus, the data at interval *t* fall outside the margin of error for the desired confidence interval and potentially a new track would be initiated.

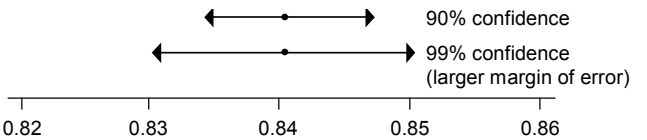

Figure 4.4 Confidence intervals of 90 and 99 percent for specimen analysis example.

If the mean change in missile position between updates was 95 m, then

$$\mu \pm z^* \sigma/\sqrt{n} = 95 \pm 2.576\ (3/\sqrt{1})\ \text{m} = 87.27\ \text{m},\ 102.73\ \text{m}. \tag{4-8}$$

Now the data at time interval t lie within the range established for 99 percent confidence.

Example 3: Suppose it is necessary to determine the center-to-center spacing of roadway sensors used for speed measurement on a section of freeway. Assume there are 25 pairs of sensors on the section, but there are resources to measure the spacing on only 3 pairs. The measurement values are 15 ft, 2.0 in (4.62 m), 15 ft, 3.0 in (4.65 m), and 14 ft, 11.0 in (4.55 m). Assume also that the standard deviation of the center-to-center sensor spacing is known from historical data to be 2.25 in (5.7 cm). What are the 90, 95, and 99 percent confidence intervals for the true center-to-center spacing of the sensors?

The sample mean of the measurements is

$$\bar{x} = (182 + 183 + 179)/3\ \text{in} = 181.3\ \text{in}\ (460.6\ \text{cm}). \tag{4-9}$$

For 90 percent confidence, $z^* = 1.645$. Thus the 90 percent confidence interval for μ is

$$\begin{aligned} \bar{x} \pm z^*\sigma/\sqrt{n} &= 181.3 \pm 1.645(2.25/\sqrt{3})\ \text{in} = 181.3 \pm 2.1\ \text{in} \\ &= 183.4\ \text{in},\ 179.2\ \text{in}\ (465.8\ \text{cm},\ 455.2\ \text{cm}). \end{aligned} \tag{4-10}$$

For 95 and 99 percent confidence, $z^* = 1.960$ and 2.576, respectively. The corresponding confidence intervals are

$$\begin{aligned} \bar{x} \pm z^*\sigma/\sqrt{n} &= 181.3 \pm 1.960(2.25/\sqrt{3})\ \text{in} = 181.3 \pm 2.5\ \text{in} \\ &= 183.8\ \text{in},\ 178.8\ \text{in}\ (466.9\ \text{cm},\ 454.2\ \text{cm}) \end{aligned} \tag{4-11}$$

for 95 percent confidence and

$$\begin{aligned} \bar{x} \pm z^*\sigma/\sqrt{n} &= 181.3 \pm 2.576(2.25/\sqrt{3})\ \text{in} = 181.3 \pm 3.3\ \text{in} \\ &= 184.6\ \text{in},\ 178.0\ \text{in}\ (468.9\ \text{cm},\ 452.1\ \text{cm}) \end{aligned} \tag{4-12}$$

for 99 percent confidence.

Thus, there is 90 percent confidence that the true center-to-center spacing lies between 179.2 in and 183.4 in (4.55 m and 4.66 m), 95 percent confidence that the true center-to-center spacing lies between 178.8 in and 183.8 in (4.54 m and 4.67 m), and 99 percent confidence that the true center-to-center spacing lies

between 178.0 in and 184.6 in (4.52 m and 4.69 m). The confidence intervals and sample mean are depicted in Figure 4.5.

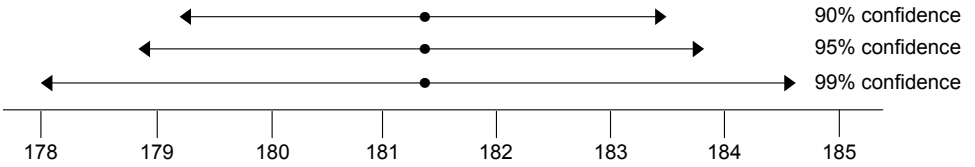

Figure 4.5 Confidence intervals of 90, 95, and 99 percent for roadway sensor spacing example.

4.4 Significance tests for hypotheses

Significance tests assess the evidence provided by data in favor of some claim about a proposition. The significance test evaluates the strength of the evidence against a postulated null hypothesis H_0, which is the statement being tested. As such, the null hypothesis is a statement of "no effect" or "no difference." The alternate hypothesis H_1 is the statement suspected to be true. Hypotheses are stated in terms of population parameters such as mean and correlation coefficient.

The probability, computed assuming H_0 is true, that the test statistic assumes a value as extreme or more extreme than that actually observed is called the *P*-value of the test. The smaller the *P*-value, the stronger is the evidence against H_0 provided by the data. If the *P*-value is as small or smaller than α, the data are said to be statistically significant at level α. That is, the data give evidence against H_0 such that H_0 occurs no more than α percent of the time. *P*-values are exact if the population distribution is normal and approximately correct for large *n* in other cases.

The *P*-value is more informative than a statement of significance because significance can now be assessed at any chosen level. For example, a result with a *P*-value equal to 0.03 is significant at the $\alpha = 0.05$ level, but not significant at the $\alpha = 0.01$ level (because $\alpha = 0.01 < P\text{-value} = 0.03$).

4.5 The *z*-test for a population mean

To test the hypothesis that μ has a specific value μ_0, the null hypothesis is constructed for H_0: $\mu = \mu_0$. The test uses the sample mean $\bar{x}$ as the population parameter and standardized variables. When the statistics are normal, the applicable standardized test statistic is the *standardized sample mean z* or *z* statistic given by

$$z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n}). \tag{4-13}$$

It is computed from a random sample of size n drawn from a population with unknown mean μ and known standard deviation σ. The z statistic has a standard normal distribution $N(\mu_0, \sigma/\sqrt{n})$ when H_0: $\mu = \mu_0$ is true.

If the alternative hypothesis is one sided on the high side, i.e., H_1: $\mu > \mu_0$, then the P-value is the probability that a standard normal random variable Z assumes a value at least as large as the observed z. In this case,

$$P = P(Z \geq z). \tag{4-14}$$

When the alternative hypothesis is one sided on the low side (i.e., the true μ is less than the hypothesized μ_0, written as H_1: $\mu < \mu_0$),

$$P = P(Z \leq z). \tag{4-15}$$

When H_1 affirms that μ is simply unequal to μ_0 (i.e., H_1 is two sided), then values of z smaller and larger than 0 count against the null hypothesis. In this case, the P-value is the probability that a standard normal random variable Z is at least as far from 0 as the observed z.

To summarize, the P-value for a test of H_0 against alternative hypotheses

$$H_1: \mu > \mu_0 \text{ is } P(Z \geq z), \tag{4-16}$$

$$H_1: \mu < \mu_0 \text{ is } P(Z \leq z), \tag{4-17}$$

and

$$H_1: \mu \neq \mu_0 \text{ is } 2P(Z \geq |z|). \tag{4-18}$$

In the double-sided test of Eq. (4-18), the probability is computed by doubling $P(Z \geq |z|)$ because the standard normal distribution is symmetric.

The following double-sided test example illustrates how the P-value is used to evaluate the truth of a hypothesis. Suppose the mean thickness of metal sheet produced by a certain process is 3 mm with a standard deviation of 0.05 mm. If the mean thickness of 5 consecutive sheets is 2.96 mm, is the process out of control?

To answer this question, set H_0: $\mu = 3$ mm and H_1: $\mu \neq 3$ mm. The P-value for testing these hypotheses is $2P(Z \geq |z|)$ calculated assuming H_0 is true. P is two sided because the sheets can be thicker or thinner than the mean.
When H_0 is true, the random variable $\bar{x}$ has a normal distribution with

$$\mu_{\bar{x}} = \mu = 3 \text{ mm} \tag{4-19}$$

and

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.05/\sqrt{5} \text{ mm} = 0.022 \text{ mm}. \tag{4-20}$$

The *P*-value is found from the normal probability calculation for the standardized sample mean $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ using a two-sided test such that

$$\begin{aligned} 2P(Z \geq |z|) &= 2P(Z \geq |(\bar{x} - \mu)/(\sigma/\sqrt{n})|) = 2P(Z \geq |(2.96 - 3)/(0.022)|) \\ &= 2P(Z \geq |1.818|) = 0.0688, \end{aligned} \tag{4-21}$$

where the probability value of 0.0688 is obtained from tables of standard normal probabilities.

Since only about 7 percent of the time will a random sample of size 5 have a mean thickness at least as far from 3 mm as that of the sample, the observed $\bar{x} = 2.96$ mm provides evidence that the process is out of control. Therefore, the null hypothesis is not confirmed.

If the sample mean was 2.98 mm, then

$$2P(Z \geq |z|) = 2P(Z \geq |(2.98 - 3)/(0.022)|) = 2P(Z \geq |0.909|) = 0.3628. \tag{4-22}$$

In this case, there is insufficient evidence to reject the null hypothesis H_0: $\mu = 3$ mm since there is a 36 percent probability that a random sample of size 5 will have a mean thickness at least as far from 3 mm as that of the sample. The result of the *P*-value calculation for $\bar{x} = 2.98$ mm is shown in Figure 4.6.

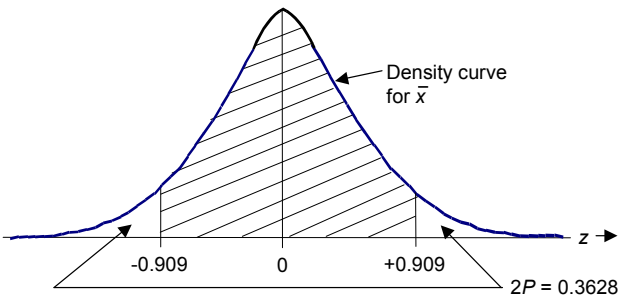

Figure 4.6 Interpretation of two-sided *P*-value for metal sheet thickness example when sample mean = 2.98 mm.

4.6 Tests with fixed significance level

Fixed significance level tests are used to decide whether evidence is statistically significant at a predetermined level without the need for calculating the *P*-value. This is accomplished by specifying a level of significance α at which a decision will occur or some other action taken. Choosing a level α in advance is appropriate if a decision has to be made, but may not be suitable if only a description of the strength of the evidence is needed. In the latter case, finding the *P*-value is more suitable.

When a fixed significance level test is appropriate, the upper p critical value z^* for the standard normal distribution is used. This value of z^* has probability

$$(1 - C)/2 = \alpha \tag{4-23}$$

to the right of it, as illustrated in Figure 4.7. If $z \geq z^*$, then the evidence is statistically significant at level α and the null hypothesis H_0 is rejected.

Values for the upper p critical value are listed in Table 4.2. Table entry for p and C is the point z^* with probability p lying above it and probability C lying between $-z^*$ and z^*. Upper p critical values were used to calculate confidence intervals in Section 4.3.

To test the hypothesis H_0: $\mu = \mu_0$ based on a random sample of size n from a population with unknown mean μ and known standard deviation σ, compute the standardized sample mean test statistic from Eq. (4-13). Then reject H_0 at a significance level α against a one-sided alternative

$$H_1: \mu > \mu_0 \text{ if } z \geq z^* \text{ or} \tag{4-24}$$

$$H_1: \mu < \mu_0 \text{ if } z \leq z^*, \tag{4-25}$$

where z^* is the upper α critical value for the standard normal distribution.

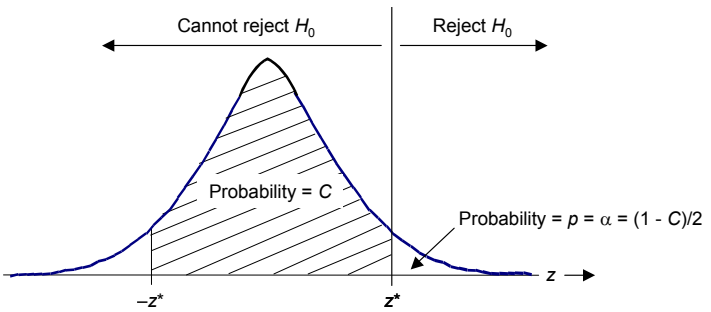

Figure 4.7 Upper critical value z^* used in fixed significance level test.

Table 4.2 Relation of upper *p* critical value and *C* to *z.**

C	*p*	*z**
50%	0.25	0.674
60%	0.20	0.841
70%	0.15	1.036
80%	0.10	1.282
90%	0.05	1.645
95%	0.025	1.960

C	*p*	*z**
96%	0.02	2.054
98%	0.01	2.326
99%	0.005	2.576
99.5%	0.0025	2.807
99.8%	0.001	3.091
99.9%	0.0005	3.291

H_0 is rejected at a significance level α against a two-sided alternative

$$H_1: \mu \neq \mu_0 \text{ if } |z| \geq z^*, \tag{4-26}$$

where z^* is now the upper $\alpha/2$ critical value for the standard normal distribution.

The two-sided alternative is evaluated using $\alpha/2$ because both the upper and lower $(1 - C)/2$ areas must be accounted for as depicted in Figure 4.8. A level α two-sided significance test rejects a hypothesis H_0: $\mu = \mu_0$ exactly when μ_0 falls outside a $(1 - \alpha)$ confidence interval for μ.

The two-sided significance test can be applied to the original metal sheet problem of Section 4.5 to evaluate whether the evidence against H_0 is statistically significant at the 10 percent level and the 1 percent level when $z = 1.818$. Since this is a two-sided test, the upper $\alpha/2$ critical value is used. Thus, $z^* = 1.645$ for $\alpha/2 = 5$ percent and $z^* = 2.576$ for $\alpha/2 = 0.5$ percent.

Since $z \geq 1.645$, the observed $\bar{x}$ provides evidence against H_0 that is significant at the 10 percent level. However, since $z < 2.576$, the observed $\bar{x}$ provides evidence against H_0 that is not significant at the 1 percent level.

An alternative way of arriving at the same conclusion is through evaluation of the confidence intervals for $C = 90$ percent and 99 percent corresponding to the $\alpha/2$ critical values illustrated in Figure 4.8. When $C = 90$ percent, $(1 - C)/2 = 0.05$ and $z^* = 1.645$ (from Table 4.2). The corresponding $(1 - \alpha)$ confidence interval, where $\alpha = (1 - C)/2$ from Eq. 4-23, is

$$\bar{x} \pm z^*\sigma/\sqrt{n} = 2.96 \pm 1.645(0.05/\sqrt{5}) \text{ mm} = 2.96 \pm 0.037 \text{ mm}$$
$$= 2.923 \text{ mm}, 2.997 \text{ mm}. \tag{4-27}$$

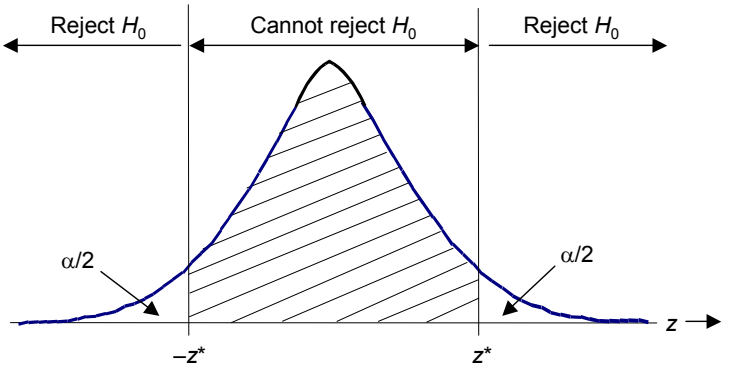

Figure 4.8 Upper and lower α/2 areas that appear in two-sided significance test.

Since the value $\mu_0 = 3$ mm falls outside this interval, the process is deemed to be out of control at the 10 percent level of significance.

When $C = 99$ percent, $(1 - C)/2 = \alpha = 0.005$ and $z^* = 2.576$. The corresponding $(1 - \alpha)$ confidence interval is

$$\begin{aligned} \bar{x} \pm z^*\sigma/\sqrt{n} &= 2.96 \pm 2.576(0.05/\sqrt{5}) \text{ mm} = 2.96 \pm 0.058 \text{ mm} \\ &= 2.902 \text{ mm}, 3.018 \text{ mm}. \end{aligned} \tag{4-28}$$

Now the value $\mu_0 = 3$ mm falls inside the confidence interval and the process is not rejected as out of control at the 1 percent level of significance.

4.7 The *t*-test for a population mean

When the standard deviation of the entire population is unknown, the standard deviation of the sample mean given by Eq. (4-1) cannot be calculated. Under these circumstances, the standard deviation s of the sample can be used in place of the standard deviation of the population. The standard deviation of the sample is calculated from the data samples x_i as

$$s = \sqrt{\frac{1}{n-1}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2]}$$

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ , \tag{4-29}$$

where n is the number of data samples drawn from the entire population and $\bar{x}$ is the sample mean. The quantity $n - 1$ represents the number of degrees of freedom, which is one less than the number of samples because the sum of the deviations $x_i - \bar{x}$ is always 0. Therefore, the last deviation can be calculated once the first $n - 1$ are known. Thus, only $n - 1$ of the squared deviations can vary freely.

When the standard deviation of the sample is substituted for the standard deviation of the entire population, the one sample t statistic given by

$$t = (\bar{x} - \mu)/(s/\sqrt{n}), \tag{4-30}$$

is substituted for the z statistic in the inference procedures discussed in Sections 4.5 and 4.6. The t statistic, denoted as $t(n - 1)$, does not have a normal distribution, but one appropriately referred to as a t distribution with $n - 1$ degrees of freedom. In terms of a random variable T having a $t(n - 1)$ distribution, the P-value for a test of H_0 against

$$H_1: \mu > \mu_0 \text{ is } P(T \geq t), \tag{4-31}$$

$$H_1: \mu < \mu_0 \text{ is } P(T \leq t), \tag{4-32}$$

and

$$H_1: \mu \neq \mu_0 \text{ is } 2P(T \geq |t|). \tag{4-33}$$

These P-values are exact if the population distribution is normal and approximately correct when n is large.

The factor $s/\sqrt{n}$ is referred to as the *standard error*. The term standard error is sometimes also applied to the standard deviation of a statistic, such as $\sigma/\sqrt{n}$ in the case of the sample mean $\bar{x}$. The estimated value $s/\sqrt{n}$ is then referred to as the estimated standard error.

The probability density curves for $t(n - 1)$ are similar in shape to the normal distribution as they are symmetric about 0 and bell shaped.[9] However, a larger amount of the area under the probability curve lies in the tails of the t distribution as shown in Figure 4.9. The tails enclose a larger area because of the added variability produced by substituting the random variable s for the fixed parameter σ. As n grows large, the $t(n - 1)$ density curve approaches the $N(0, 1)$ curve more closely since s approaches σ as the sample size increases.

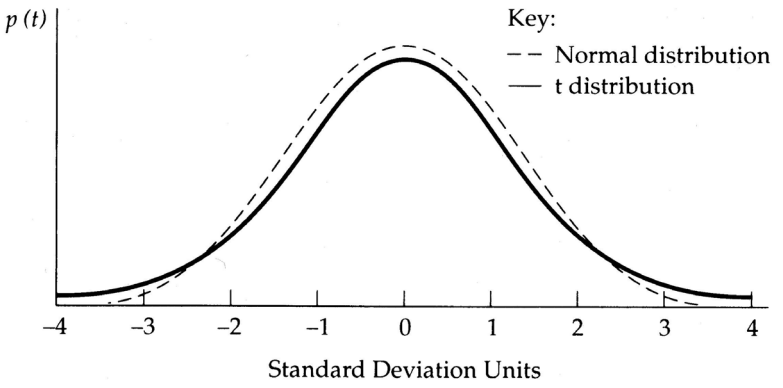

Figure 4.9 Comparison of *t* distribution with four degrees of freedom with standardized normal distribution. (D. Knoke and G.W. Bohrnstedt, *Basic Social Statistics*, Itasca, IL: F.E. Peacock Publishers [1991].)

When the standard deviation of the sample mean is substituted for the standard deviation of the population, a level C confidence interval for μ is computed using t^* as

$$\bar{x} \pm t^* s/\sqrt{n} ,$$

where t^* is the upper $(1 - C)/2$ critical value for the $t(n - 1)$ distribution.

Table 4.3 contains values of t^* for several confidence levels (i.e., the probability enclosed by the central area of the t distribution) as a function of the number of degrees of freedom. When the number of degrees of freedom is infinite, t^* equals z^* as illustrated for ∞ degrees of freedom in Table 4.3. The entries in this row are equal to those in Tables 4.1 and 4.2.

To illustrate the effect on the confidence interval of substituting the standard deviation of the sample mean for the standard deviation of the entire population, the 90 and 99 percent confidence intervals for the specimen analysis example in Section 4.3 is recomputed. The standard deviation of the specimen samples is found from Eq. (4-29) as

$$s = 0.0042 \text{ g/l}. \tag{4-34}$$

The 90 percent confidence interval for μ is

$$\bar{x} \pm t^* s/\sqrt{n} = 0.8404 \pm 2.920\ (0.0042/\sqrt{3}\ ) \text{ g/l} = 0.8404 \pm 0.0071 \text{ g/l}$$
$$= 0.8333 \text{ g/l}, 0.8475 \text{ g/l}. \tag{4-35}$$

Table 4.3 Values of *t for several confidence levels and degrees of freedom.**

Degrees of Freedom	(1 – *C*)/2							
	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
5	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
10	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
15	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
20	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
30	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
∞	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence Level							

The 99 percent confidence interval for μ is

$$\bar{x} \pm t^* \, s/\sqrt{n} = 0.8404 \pm 9.925 \, (0.0042/\sqrt{3}) \text{ g/l} = 0.8404 \pm 0.0241 \text{ g/l}$$
$$= 0.8163 \text{ g/l}, 0.8645 \text{ g/l}. \quad (4\text{-}36)$$

Table 4.4 compares the confidence intervals from the *z*- and *t*-tests. As expected, the confidence intervals at each confidence level are larger when the standard error and *t** are used.

Table 4.4 Comparison of *z*-test and *t*-test confidence intervals.

Confidence Level	*z*-test Confidence Interval	*t*-test Confidence Interval
90%	0.8339 to 0.8469	0.8333 to 0.8475
99%	0.8303 to 0.8505	0.8163 to 0.8645

4.8 Caution in use of significance tests

When a null hypothesis can be rejected at low values of α (e.g., 0.05 or 0.01), there is good evidence that an effect is present, but that effect may be extremely small. Thus, the low significance level does not mean that there is strong association, only that there is strong evidence of some association.

Significance tests and confidence intervals are based on laws of probability. Therefore, randomization in sampling or experimentation ensures that randomized samples are obtained and that these laws apply. There is no way to make data into simple random samples if they are not gathered as such in the first place. Analyzing data that are not from simple random samples will not produce valid inferences even if the above statistical techniques are used. Data must be examined for outliers and other deviations from a consistent pattern that would cause the samples to be suspect.

4.9 Inference as a decision

Statistical inference provides answers to specific questions, along with a statement of the confidence in the correctness of the answer. A level of significance α chosen in advance points to the outcome of the test as a decision. Accordingly, if the P-value is less than α, reject H_0 in favor of H_1. Otherwise, do not reject H_0. The transition from measuring the strength of evidence to making a decision is not a small step. A decision should be reached only after the evidence from many studies or data acquisition periods or sources is weighted.[8]

When inference methods are used for decision making, the null hypothesis is no longer singled out as a special type of outcome (as it is in significance testing). In decision making there are simply two hypotheses from which we must select one and reject the other. Hypothesis H_0 no longer enjoys special status as the null hypothesis.

The significance level, like the confidence level, gives information about how reliable the test method is in repeated use. Thus, if 5 percent significance tests are repeatedly used to evaluate the truth of H_0 when H_0 is in fact true, a wrong decision will be reached 5 percent of the time (i.e., the test will reject H_0) and a correct decision reached 95 percent of the time (i.e., the test will fail to reject H_0).

High confidence is of little value if the confidence interval is so wide that few values of the parameter are excluded. Thus, a test with small α almost never rejects H_0 even when the true parameter value is far from the hypothesized value. A useful test must be able to detect that H_0 is false as well as be concerned about the margin of error of a confidence interval. The ability of a test to satisfy the latter concerns is measured by the probability that the test will reject H_0 when an alternative is true. As this probability increases, so does the sensitivity of the test. The probability that the test will reject H_0 is different for different values of the parameter associated with the alternate hypothesis H_1. As described below, this probability is related to the power of the test. Qualitatively, the power of a test is the probability that the test will detect an effect of the size hoped for.

In light of the above discussion, a wrong decision is reached when one of two types of errors occurs. These are the Type 1 and Type 2 errors, which were

depicted in the classical inference concept illustrated in Figure 3.6. A Type 1 error rejects H_0 and accepts H_1 when in fact H_0 is true. A Type 2 error accepts H_0 and rejects H_1 when in fact H_1 is true. The two correct and two incorrect situations arising in hypothesis testing are summarized in Table 4.5. The probabilities of their occurrence are also shown.

Table 4.5 Type 1 and Type 2 errors in decision making.

	Truth about the population (*True state of nature*)	
Decision	$\boldsymbol{H_0}$ **True**	$\boldsymbol{H_1}$ **True**
Reject $\boldsymbol{H_0}$	Type 1 error Probability = α	Correct decision Probability = $1 - \beta$
Accept $\boldsymbol{H_0}$	Correct decision Probability = $1 - \alpha$	Type 2 error Probability = β

Type 1 and Type 2 error value selection is dependent on the consequences of a wrong decision, e.g., is the application one of missile interception, aircraft identification, commercial vehicle classification, or historical data collection?

The significance level α of any fixed level test is the probability of a Type 1 error. Thus α is the probability that the test will reject hypothesis H_0 when H_0 is in fact true. The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the *power of the test against that alternative*. The power is equal to 1 minus the probability of a Type 2 error for that alternative. If the Type 2 error is denoted by β, the power of a test for that alternative is given by $1 - \beta$.

High power is desirable. The numerical value of the power is dependent on the particular parameter value chosen in H_1. For example, values of the mean μ that are in H_1 but lie close to the hypothesized value μ_0 are harder to detect (lower power) than values of μ that are far from μ_0. Using a significance test with low power makes it unlikely to find a significant effect even if the truth is far from hypothesis H_0. A hypothesis H_0 that is in fact false can become widely believed if repeated attempts to find evidence against it fail because of low power.

Two examples are cited below to show how the power of a test is calculated and what inferences can be drawn from each result.

Single-sided power of a test example: suppose a cheesemaker determines that the milk from one producer is heavily watered from measurements of its freezing point.[8] Five lots of milk are sampled and the freezing points of each are measured. The mean freezing point determined from the five samples is $\bar{x} = -0.539^{o}$C, whereas the mean freezing temperature of milk is normally -0.545^{o}C with a standard deviation of $\sigma = 0.008^{o}$C. Furthermore, suppose the

cheesemaker determines that milk with a freezing point of –0.53°C will damage the quality of his cheese. Will a 5 percent significance test of the hypothesis

$$H_0: \mu \geq -0.545^{\circ}\text{C}$$

based on the sample of five lots usually detect a mean freezing point this high?

The question can be answered by finding the power of the test against the specific alternative $\mu = -0.53^{\circ}$C.

The test measures the freezing point of five lots of milk from a producer and rejects H_0 when

$$z = [\bar{x} - (-0.545)]/(0.008/\sqrt{5}) \geq 1.645, \tag{4-37}$$

where 1.645 is the upper p critical value for $\alpha = 5$ percent.

This is equivalent to

$$\bar{x} \geq -0.545 + (1.645)(0.008/\sqrt{5}) = -0.539^{\circ}\text{C}. \tag{4-38}$$

Since the significance level is $\alpha = 0.05$, this event has a probability of 0.05 occurrance when in fact the population mean μ is -0.545°C. The notation expressing that the probability calculation assumes $\mu = -0.545^{\circ}$C is

$$P(\bar{x} \geq -0.539 | \mu = -0.545) = 0.05. \tag{4-39}$$

The power of the test against the alternative -0.53°C is the probability that H_0 will be rejected when in fact $\mu = -0.53^{\circ}$C, which is written as

$$P(\bar{x} \geq -0.539 | \mu = -0.53). \tag{4-40}$$

The probability in Eq. (4-40) is calculated by standardizing $\bar{x}$ using the value $\mu = -0.53$ for the population mean and the original value of 0.008 for the population standard deviation. Thus,

$$\begin{aligned} P(\bar{x} \geq -0.539 | \mu = -0.53) &= P\{[\bar{x} - (-0.53)]/(0.008/\sqrt{5}) \\ &\geq [-0.539 - (-0.53)]/(0.008/\sqrt{5})\} \\ &= P(Z > -2.52) = 0.9941. \end{aligned} \tag{4-41}$$

The power of the test for the sampling distribution $\bar{x}$ when $\mu = -0.53^{\circ}$C is true is shown in Figure 4.10. This significance test is sensitive enough for the

cheesemaker's application since it will almost always (with probability greater than 99 percent) reject H_0 when in fact $\mu = -0.53^{o}C$.

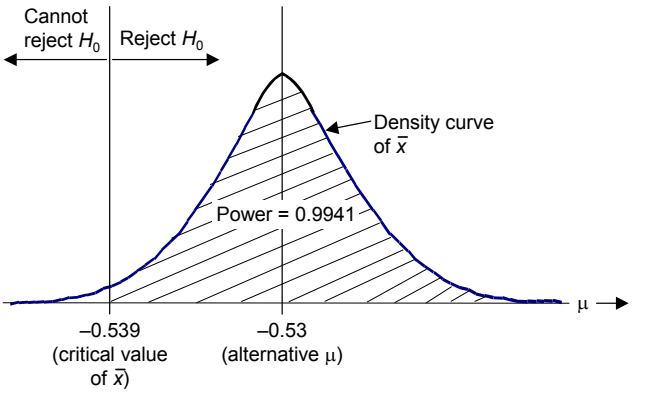

Figure 4.10 Hypothesis rejection regions for single-sided power of a test example.

Double-sided power of a test example: the double-sided power of a test calculation is illustrated by referring to the metal sheet example described in Section 4.5. The power of the test against the specific alternative $\mu = 2.97$ mm is found below.

The hypothesis H_0 was rejected in the original example ($\mu = 3$ mm, $\bar{x} = 2.96$ mm) at the 10 percent level of significance or when $z^* = 1.645$ since P was 0.0688 or less than 10 percent. Equivalently, the test rejects H_0 when either of the following is true:

(1) $z \geq 1.645$ or equivalently when $\bar{x} \geq 3.036$, where z and $\bar{x}$ are related by

$$z = (\bar{x} - \mu)/(\sigma/\sqrt{n}) = (\bar{x} - 3)/0.022 \tag{4-42}$$

or

(2) $z \leq -1.645$ or $\bar{x} \leq 2.964$.

Since these are disjoint events, the power is the sum of their probabilities computed assuming the alternative $\mu = 2.97$ mm is true. Thus,

$$\begin{aligned} P(\bar{x} \geq 3.036|\mu = 2.97) &= P[(\bar{x} - 2.97)/0.022 \geq (3.036 - 2.97)/0.022] \\ &= P(Z \geq 3.00) = 0.0013, \end{aligned} \tag{4-43}$$

and

$$P(\bar{x} \le 2.964|\mu = 2.97) = P[(\bar{x} - 2.97)/0.022 \le (2.964 - 2.97)/0.022] = P(Z \le 0.273) = 0.606. \quad (4\text{-}44)$$

Since the power is approximately 0.607, we cannot be confident that the test will reject H_0 when the alternative is true. This situation is shown in Figure 4.11. If the power was greater than 0.9, then we could be quite confident that the test would reject H_0 when the alternative is true.

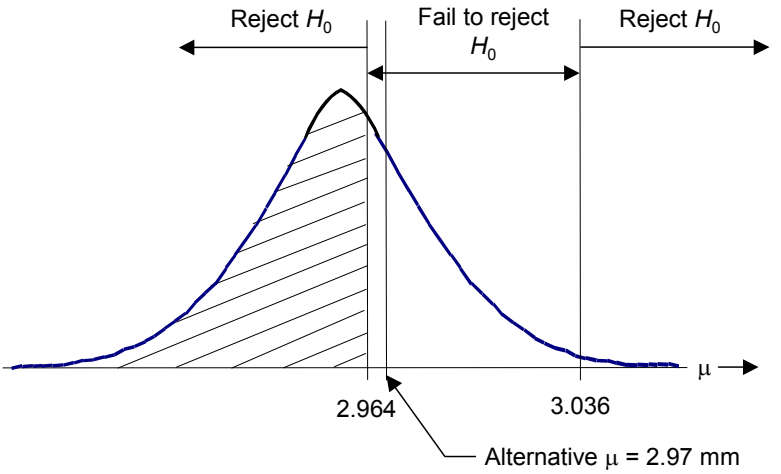

Figure 4.11 Hypothesis rejection regions for double-sided power of a test example.

4.10 Summary

Data distributions are defined by statistics such as expected values, standard deviations, and shape parameters. The sample mean $\bar{x}$ is an unbiased estimator of an unknown population mean μ if the samples are randomly obtained and are representative of the entire population. The standard deviation of the sample mean is calculated by dividing the standard deviation of the population by the square root of the number of observations in the sample. Confidence levels express a probability C that a sample from any normal distribution falls within $\pm z^*$ standard deviations of the distribution's mean. A level C confidence interval for a parameter is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter. The value z^* for confidence C encompasses the central area C between $-z^*$ and z^*.

Significance tests assess the evidence provided by data in favor of some claim about a proposition. When significance tests are used, the null hypothesis H_0 is the statement being tested. The significance test is designed to assess the strength of the evidence against the null hypothesis. The alternate hypothesis H_1 is the statement suspected of being true. The probability, computed assuming H_0 is true, that the test statistic assumes a value as extreme or more extreme than that actually observed is called the *P*-value of the test. The smaller the *P*-value, the stronger is the evidence against H_0 provided by the data. If the *P*-value is as small or smaller than α, the data are said to be statistically significant at level α. Single- and double-sided hypothesis tests that compare the probability of a sample parameter having a specific value are performed using a test statistic such as the *standardized sample mean z* or *z* statistic. The *z* statistic has a standard normal distribution $N(\mu_0, \sigma/\sqrt{n})$ when H_0: $\mu = \mu_0$ is true. Fixed significance level tests are used to decide whether evidence is statistically significant at a predetermined level without the need for calculating the *P*-value. This is accomplished by specifying, in advance, a level of significance α at which a decision will occur or some other action taken.

When the standard deviation of the entire population is unknown, the standard deviation *s* of the sample can be used in place of the standard deviation of the population to calculate an estimate for the standard error of the sample mean. When *s* is used, the *t* statistic replaces the *z* statistic in inference procedures and *t** replaces *z** when calculating confidence intervals.

When inference methods are used for decision making, the null hypothesis is no longer singled out as a special type of outcome (as it is in significance testing). In decision making there are simply two hypotheses from which one is selected and the other rejected. A decision may be wrong, however, due to two types of errors, Type 1 and Type 2. A Type 1 error rejects H_0 and accepts H_1 when in fact H_0 is true. A Type 2 error accepts H_0 and rejects H_1 when in fact H_1 is true.

Classical inference procedures cannot be applied when data are haphazardly collected with bias of unknown size. Since the sample mean is not resistant to outliers, outliers can have a large effect on the confidence interval. Therefore, outliers should be identified and their removal justified before computing a confidence interval. If the outliers cannot be removed, procedures should be found that are insensitive to outliers. If the sample size is small and the population is not normal, the true confidence level will be different from the value *C* used in computing the interval. Sensitivity to nonnormal populations is not large when $n \geq 15$ in the absence of extreme outliers and skewness.

Table 4.6 summarizes the strengths and weaknesses of classical inference.

Table 4.6 Characteristics of classical inference method.

Strengths	Weaknesses
Probability model links observed data and a population	When generalized to include multi-dimensional data from multiple sensors, *a priori* knowledge and multidimensional probability density functions are required
Probability model is usually empirically based on parameters calculated from a large number of samples	Generally, only two hypotheses can be assessed at a time, namely H_0 and H_1
A number of decision rules may be used to decide between the null hypothesis H_0 and an opposing hypothesis H_1	Multivariate data produce evaluation complexities *A priori* assessments are not used

References

1. W.G. Cochran, *Sampling Techniques*, 3rd Ed., John Wiley and Sons, New York, NY (1977).
2. R.M. Groves, *Survey Errors and Survey Costs*, John Wiley and Sons, New York, NY (1989).
3. L. Kish, *Statistical Design for Research*, John Wiley and Sons, New York, NY (1987).
4. J.T. Lessler and W.D. Kalsbeek, *Nonsampling Error in Surveys*, John Wiley and Sons, New York, NY (1992).
5. P. Levy and S. Lemeshow, *Sampling of Populations: Methods and Applications*, 2nd Ed., John Wiley and Sons, New York, NY (1991).
6. D. Raj, *The Design of Sample Surveys*, McGraw Hill, New York, NY (1972).
7. L.L. Chao, *Statistics: Methods and Analysis*, McGraw-Hill, New York, NY (1969).
8. D.S. Moore and G.P. McCabe, *Introduction to the Practice of Statistics*, 4th Ed., W.H. Freeman and Company, New York, NY (Aug. 2002).
9. D. Knoke and G.W. Bohrnstedt, *Basic Social Statistics*, F.E. Peacock Publishers, Itasca, IL (1991).