Chapter 4

Confidence Interval of the Mean

4.1 The Basics

Managers should base their decisions from data analysis using estimated, stable population values that persist into the future. This chapter introduces *statistical inference*, the estimation of these population values. The focus here is the estimation of the population mean, μ.

statistical inference: Estimation of a population value from analysis of data sampled from that population.

4.1.1 Why Do We Care?

Management decisions impact the future. A sample statistic, unfortunately, depends not only upon a stable, underlying value of interest, but also upon inherently unstable sample values influenced by random error. Obtain the best forecast of future values of a variable from knowledge of the true underlying reality, the relevant stable population value for the variable of interest. The difficulty is that population values are abstractions, which cannot be observed or directly measured. Fortunately, the past provides the information from which to estimate a population value to forecast the future.

This understanding of reality begins with the data summaries provided by *descriptive statistics*. Previous chapters demonstrated how to describe and summarize the data. Here our focus shifts to inferential statistics – from merely summarizing data, such as the sample mean, to estimating the true reality that underlies the entire population from which the data were sampled, such as the population mean. We bring a fresh point of view to the same data. We now *infer* the unknown value of the population mean, μ, common not only to all the data values of a variable in the sample but to *all* the past, present, and future values for the process that generates the data values.

descriptive statistics: Summaries of sample data.

The estimate of the population mean, μ, begins with the sample of data and its mean, m. The problem is that a sample mean generally does not equal the corresponding population mean. More useful is the knowledge of how close the estimate is likely to be to the corresponding population value. A useful form of the estimate specifies a range of values that likely contains the true population mean.

key concept: The estimate m from the data provides useful information for decision making only if there is confidence that m is close to μ.

For example, consider the clean-up cost for a surgical procedure at a given hospital. The procedure has been implemented many times, and each time the cost assessed.

Suppose the sample mean of the cost of these procedures is $m = \$2984$. Without assessing confidence, this sample statistic yields a data summary of unknown quality in relation to the true mean, μ. Only by an improbable coincidence would μ exactly equal $m = \$2984$. So reframe the question. How close is m likely to be to μ?

confidence interval: A *range of values* that likely contains the population value.

Consider statistical inference in the form of the *confidence interval.* This form of an estimate of μ replaces the corresponding statistic, here the sample mean, m, with a range of values about the sample statistic, the *confidence interval.* The width of the confidence interval depends, in part, on the chosen *confidence level*, the degree of confidence that the interval contains the population value. This number is usually close to 100%, such as the commonly chosen confidence level of 95%.

confidence level: Degree of confidence, such as 95%, that the interval contains the population value of interest.

If the value of μ is likely within the interval $\$2984 \pm \$845$, the confidence interval, then there is *no* actionable information. In contrast, if the value of μ is likely within $\$2984 \pm \$8.45$, then this same sample mean within this much smaller interval provides *much* information. Always provide a confidence interval instead of a sample estimate by itself. That is, always follow the analysis of descriptive statistics with the analysis of inferential statistics.

key concept: Begin data analysis with descriptive statistics, end with inferential statistics.

4.1.2 Brief Application

Consider again the cost of the clean-up of a type of surgery. The cleanup costs of 100 different procedures conducted at the same hospital are the data values for a single variable called CleanUp. Find the data on the web in a file named SurgeryCost.csv.

```
http://lessRstats.com/data/SurgeryCost.csv
```

The data table contains one column of the 100 data values listed underneath the variable name CleanUp. Figure 4.1 lists the first few and last few lines of the data file. Examine the full data file by pointing a web browser to its URL (web address).

```
CleanUp
3071.71
2994.08
2922.17
3063.56
. . .
3075.34
3087.72
```

Listing 4.1: First and last data values for the variable CleanUp in the data file SurgeryCost.csv.

To analyze the data with `R`, first read the data from the web file SurgeryCost.csv into a data table within `R` called `d`. Use the `lessR` function `Read()` to read the data into the `R` data table, called a data frame.

```
d <- Read("http://lessRstats.com/data/SurgeryCost.csv")
```

The `<-` indicates to place that data values read from the file directly into the R data table called *d*.

Obtain the confidence interval of the mean for the variable named CleanUp with the `lessR` function `ttest()`, abbreviated `tt()`. As with other `lessR` analysis functions, by default the analysis presumes the data values for the variables of interest are in the *d* data table. For simplicity, invoke the brief version of the `ttest()` analysis with `tt_brief()` to calculate the confidence interval from the data values.

```
tt_brief(CleanUp)
```

The primary output of `tt_brief()`, shown in Listing 4.2, includes the summary statistics in the first line and the 95% confidence interval in the last line.

```
CleanUp:  n.miss = 0,  n = 100,   mean = 2984.170,  sd = 42.590

Margin of Error for 95% Confidence Level:  8.451
95% Confidence Interval for Mean:  2975.719 to 2992.621
```

Listing 4.2: Primary output of `tt_brief()` for the analysis of the CleanUp data.

The output in Listing 4.2 for the variable CleanUp provides the sample size, sample mean, and sample standard deviation as $n = 100$, $m = \$2984.17$ and $s = \$42.59$. The data values are not available in some situations, but the summary statistics are known from some previous analysis. In this situation, obtain the identical output from the following function call to `ttest()` that references the summary statistics, but not the data values.

```
tt_brief(n=100, m=2984.17, s=42.59)
```

Regardless of whether the analysis was obtained directly from the data values or the resulting summary statistics of the data, the computation of the margin of error for this 95% confidence interval yields $8.45. The corresponding confidence interval is $m \pm$ the margin of error. Specify the confidence interval by its endpoints, its lower and upper bounds.

margin of error: The maximum expected difference between the true population value and a sample estimate of that value.

Lower Bound: $\$2984.170 - \$8.451 = \$2975.72$
Upper Bound: $\$2984.170 + \$8.451 = \$2992.62$

Interpret the margin of error and associated confidence interval in the context of the chosen level of confidence, here the default value of 95%. Section 4.3.2 illustrates the computation of this margin of error and confidence interval.

`margin of error`, Section 4.3.2, p. 91

Interpret the confidence interval as follows.

> With 95% confidence, the true average cost of the surgery clean-up for this hospital is somewhere between $2976 and $2993.

Purpose	Interpretation
State the Confidence Level	With 95% confidence
Unknown population mean μ	the true average
Variable of interest	cost of the surgery clean-up
Population where sample is obtained	for this hospital
CI is a range, not a single value	is somewhere between
Lower and Upper bounds of CI	\$2976 and \$2993.

Table 4.1: Template for the interpretation of a confidence interval.

The template in Table 4.1 shows how to construct this interpretation.

The primary technical term in the interpretation of the confidence interval is the phrase "95% confidence". What does this phrase mean? As is true of all inferential concepts derived from classical statistics, the concept derives meaning over usually hypothetical repeated sampling of samples of the same size from the same population. Of course, in practice, only one sample is typically observed. Still, formulas from the mathematicians allow us to proceed *as if* we had multiple samples from information derived only from a single sample.

meaning of 95% confidence level: 95% of many, many repeated samples will contain the corresponding population value.

The issue is that each random sample from the same population yields a *different* sample mean and margin of error, so each random sample yields a *different* confidence interval. There is nothing special about the specific lower and upper bounds \$2975.72 and \$2992.62 of this one obtained confidence interval from this one sample of data. Another random sample of data would result in a different confidence interval, with different lower and upper bounds and even a different width. Instead, to assert "95% confidence" means that for every 100 samples obtained and 100 corresponding unique confidence intervals computed, on average, 95 of the intervals will contain the unknown value of μ, and 5 will not.

The true value of μ, here the population mean for the process of post-surgery clean-up, may or may not be in any one confidence interval calculated from sample data. With the 95% confidence interval we are 95% confident that this one obtained confidence interval, here \$2975.72 to \$2992.62, does contain the true mean. Still, as with the outcome of any random process, we never know for sure.

As we see, unfortunately, the desired population value is not guaranteed to lie within the interval. For a given set of a variable's data values, the only way to be more confident that the confidence interval contains the desired population value is to construct a wider interval. A confidence interval at a 99% confidence level provides more confidence that the population value lies within the interval than does a 95% level, but unfortunately only at the expense of a wider interval. For most applications, 95% represents a reasonable trade-off between precision and amount of confidence.

Confidence Level, Section 4.5.2, p. 99

4.2 Conceptual Basis of the Confidence Interval

The confidence interval as a primary form of statistical inference is a concept central to data analysis. What are the underlying concepts and the motivations upon which the confidence interval is based? The basis for the confidence interval is explored next.

4.2.1 The Sample Mean, m, as a Variable

We begin with an analysis of a population with reasonably well understood characteristics, such as for the variable Height of adult USA women. The values of height conform to an approximate normally distribution. The population mean is about $\mu = 65.5$ in, the population standard deviation is about $\sigma = 2.5$ in. Figure 4.1 illustrates this normal distribution of heights, demarcated in units of the standard deviation of 2.5.

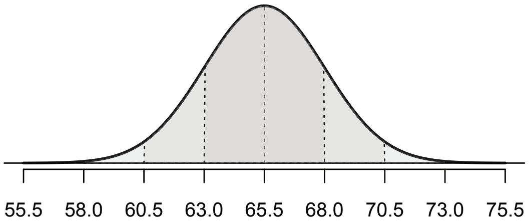


Figure 4.1: Distribution of all values from a normal population with a mean of $\mu = 65.5$ and a standard deviation of $\sigma = 2.5$.

What are the characteristics of a random sample of data drawn from this population? Every random sample yields different values, but the type of data values likely to be observed in such a sample are constrained by the population mean and standard deviation. Figure 4.1 shows that the most frequently occurring heights are close to the value of the population mean, 65.5 in. Further, not many values are more than three standard deviations or $(3)(2.5) = 7.5$ inches from the mean. That is, relatively few women are taller than 73 in or shorter than than 58 in.

To obtain a sample of randomly drawn data values from this population, suppose the height of each of the millions of American women was written on a small piece of paper and all these pieces of paper placed in a very large bag. Then shuffle the contents of the bag and, say, blindly pull 10 of these pieces of paper from the bag. The result is a random sample of $n = 10$ of the measurements of Height, amenable for data analysis.

Fortunately, we can emulate this sampling procedure with the computer, which can simulate the effect of drawing a random sample. We can invoke a function that will return results *as if* we had actually had the heights of all adult American women and then generate a random sample of a specified size from that population. One way to accomplish this task is with the `lessR` function `simMeans()`. Specify the number of samples with parameter `ns`, the size of each sample with `n`, and the population values with `mu` and `sigma`.

```
simMeans(ns=2, n=10, mu=65.5, sigma=2.5)
```

The output of the first sample of 10 random data values follows in Listing 4.3.

```
Sample  Mean   SD     1     2     3     4     5     6     7     8     9    10
     1  64.6  1.1  65.4  66.0  62.5  65.8  64.3  63.6  65.4  64.7  64.5  64.2
```

Listing 4.3: First set of 10 simulated data values for women's height.

The mean for this sample is $m = 64.6$, a little under the true value of $\mu = 65.5$. But what if we were in the typical data analysis situation and did not already know the value of μ? What if all that we knew regarding women's heights was limited to these 10 data values and corresponding statistics such as $m = 64.5$. What would we estimate as the value of μ from this information? Although μ could be practically anything, the sample value m is likely close to its corresponding population value, μ. How close? Unfortunately, just to know the value of m for a single sample for a variable does not provide enough information to know how close any one value of m likely is to the unknown value of μ.

The analysis of the data values of a variable typically is based on only a single sample. The logic of statistical inference, however, follows from what occurs over multiple samples of the same size, n, all drawn from the same population of values for the variable. Because the sampling process is random, every new sample yields a new set of data values and thus a new value of the sample mean, m. This variation of the value of m from sample to sample is an example of *sampling variation*.

sampling variation: The value of a statistic varies from one sample to another (usually hypothetical) sample.

To illustrate sampling variation, find the second random sample drawn from the population of women's heights with the `simMeans()` function in Listing 4.4, which yields a different set of data values.

```
Sample  Mean   SD     1     2     3     4     5     6     7     8     9    10
     2  65.2  3.1  67.8  65.2  65.8  59.2  65.9  67.8  69.1  63.6  61.1  66.9
```

Listing 4.4: Second set of 10 simulated data values for women's height.

For this second sample, the value of the sample mean is $m = 65.2$, a value not equal to the true population mean of $\mu = 65.5$, nor the value of the first sample mean, $m = 64.6$. This conceptual insight is the abstraction that provides the central framework for statistical inference. The values of the variable m vary across repeated, albeit usually hypothetical, samples, each of size n. To estimate the constant value of μ requires the analysis of *two* variables: the sampled variable of interest, generically referred to as Y, and m, defined over multiple samples of the population from which the actual sample was drawn.

Fortunately, mathematicians have developed expressions that use information obtained from only one sample of data values to estimate what *would* occur regarding this sampling variation *if* we had multiple samples. We analyze a single sample of data, but the conceptual framework that underlies this analysis describes an indefinitely large set of similar randomly obtained samples. Such are the wonders of mathematics.

4.2.2 The Standard Error

As discussed, a statistic calculated from what is usually *one* obtained sample, such as the sample mean, m, is useful for decision making only if it is close to its true corresponding population value, here μ. Yet m varies from sample to sample, so how to evaluate if the one observed m is likely close to μ? A key realization is that the smaller the range of variability of m from sample to sample, the more likely that any one m is close to μ.

key concept: The less m varies across samples, the more likely any one m is close to μ.

Consider again the sample of the variable Height, of which 10 values yielded $m = 64.6$. How to estimate the value of μ from the data? Hypothetically, consider an additional 7 samples, each also of $n = 10$. From these eight samples, obtain eight different values of m, which suppose range from a minimum of 64.1 to a maximum of 65.2. Given this information, what is the value of μ?

One answer is that the exact value of μ cannot be known from the 8 different obtained values of m. A more useful answer is that the extent of the sampling variability of m provides a guide as to the value of μ, which is *likely* somewhere between the minimum and maximum obtained values of m, here between 64.1 and 65.2. This range of values of m is an informal example of a confidence interval, which likely contains the *unknown* value of μ.

In contrast, consider a highly variable set of m's over multiple samples from a population. Suppose the values of m range from 30 all the way to 100. In this situation, all we would know is that the true mean, μ, is likely to be somewhere between 30 and 100. In this situation, knowing any one m provides little confidence as to the value of μ.

The variability of many m's over many samples indicates how close any one m is likely to be to μ. The less variability the better the quality of our estimation in general. To obtain the confidence interval to estimate μ, we need a method to assess the variability of the sample means, the m's over repeated samples. Presumably, we would know this range of variation without actually having to obtain these additional samples.

As with any variable, m has a population mean and standard deviation. Assess the variability of m over multiple samples with its standard deviation, what is called the *standard error* of the sample mean. The standard error is a standard deviation of a statistic over (usually hypothetical) samples. The reference to the "standard deviation" of a variable usually references the variability of the data values.

standard error: The standard deviation of a *statistic* across (usually hypothetical) multiple samples.

Mathematicians have discovered the relationship that allows the computation of the mean's standard error without having to gather multiple samples. We begin with the expression for the standard error expressed in terms of population values. Denote the population standard deviation of m, its standard error defined over *all* possible samples of size n, with σ_m.

standard error of the sample mean: Standard deviation of all possible data values divided by the square root of the size of each sample.

Express this standard error with a simple expression in terms of the standard deviation of the values of this population and the size of each sample, n, randomly sampled from this population.

Actual (population) standard error of m: $\quad \sigma_m = \dfrac{\sigma}{\sqrt{n}}$

The population standard deviation, σ, is the standard deviation of the Y, the population from which the data sample was obtained. This population standard deviation is typically unknown, so the σ_m cannot be directly computed. This discussion, however, provides the logical basis for what occurs with actual data analysis.

Both variables, the variable of interest such as Height, generically referred to as Y, and the sample mean, m, of the data values of the variable of interest, have the same population mean, μ. That is, the population mean of the values for Y is the same as the mean of the means of all the samples from the population of Y. The distributions of Y, and of m defined over repeated samples of Y, are centered over the same value, μ.

Distinguish the distributions of the two variables Y and m by their variabilities. The standard deviation of m is the standard deviation of the values of the population from which the data values were sampled, divided by the square root of the sample size of all the samples, n. Because n is a positive number, the sample mean, m, varies less than the values of the variable Y from which it is computed. The standard deviation of the sample mean, m, its standard error, is smaller than the standard deviation of the data from which the mean is computed.

key concept: The sample mean (across samples) varies less than the corresponding data.

That the standard error is smaller than the standard deviation is clear from the formula, but why is this? The process of computing the mean is a centering process: The large values in a sample tend to cancel out the small values in the same sample. A sample of Heights, for example, may have some large Heights and some small Heights, but when averaged the large and small values cancel each other out in the computation of the mean. The result is that the sample mean fluctuates more closely around the population mean μ than do individual data values.

4.2.3 The Probability Interval

Central Limit Theorem: m is at least approximately normal except for small samples from skewed populations.

`Appendix: CLT`, Section 4.6, p. 178

If the values of Y in the population are normally distributed, then so is the corresponding (usually hypothetical) distribution of the sample mean, m, for any sample size. But there is more. Mathematicians have discovered a remarkable result: *Regardless* of how the original variable Y is distributed, normal or not, for an adequate sample size, n, the corresponding variable m is at least approximately normally distributed. This result applies the *central limit theorem* (CLT), explored in more detail in the appendix. An “adequate” sample size usually means at least $n = 30$, and even less if the distribution of the variable Y is not skewed.

`normal curve probabilities`, Section 3.2, p. 55

As we have learned, the probabilities for a specified range of a normally distributed variable follow from the value of the variable’s standard deviation. And we also know, at least theoretically, the standard deviation of m, its standard error, given the standard deviation of the values of the population from which the data were sampled and the size of the corresponding sample. The result is that normal curve probabilities, in conjunction with the standard error of m, permit us to calculate the range of variation of m.

CLT [practical implication]: If $n > 30$, assume m is normal, otherwise inspect the data.

The practical implication of the Central Limit Theorem is that before conducting an inferential analysis to estimate μ, first verify that m is (at least approximately) normally distributed, so that normal probabilities apply to the distribution of m.

For Data Analysis: Apply the Central Limit Theorem

- If the sample size, n, is larger than 30 or so, then assume m is normal.
- If the sample size, n, is much less than 30, inspect a histogram of the data to ascertain that skewness is not too severe.

z-values and normal curve probabilities, Section 3.2, p. 55

Obtain the range of the sampling variability of m from the knowledge of its presumed normality and standard error, its standard deviation. We know that 95% of all the values of a normal distribution lie within 1.96 standard deviations of its population mean. That is, 95% of all z-values of the normally distributed variable, here m, are between -1.96 and 1.96.

$$z\text{-value of the sample mean: } z_m = \frac{m - \mu}{\sigma_m}$$

cutoff value: Value of a distribution that cut offs a specified percentage of values.

The *cutoff value* of 1.96 cuts off the upper 2.5% of the standardized values of a normally distributed variable, and -1.96 cuts off the bottom 2.5% of the values, leaving 95% of the values in between. Denote these cutoff values accordingly as $z_{0.25}$ and $-z_{0.25}$. Define this 95% range of variability about μ between $-z_{0.25}$ and $z_{0.25}$ as the 95% probability interval of the standardized value of m, as illustrated in Figure 4.2.

probability interval: Range of variation that contains a specified percentage of values.

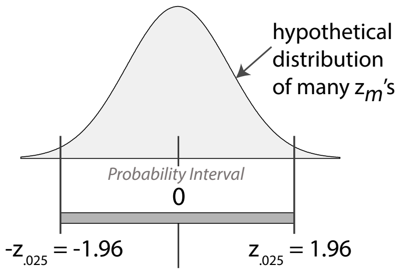


Figure 4.2: Theoretical 95% probability interval that contains 95% of all possible standardized sample means, z_m, calculated for samples each of size n for the variable of interest.

Instead of standardized values, express the probability interval with the units used to measure the variable of interest. Instead of the distribution of the standardized means about 0, express the probability interval in terms of the distribution of the means in their unit of measurement, centered about μ, as shown in Figure 4.3.

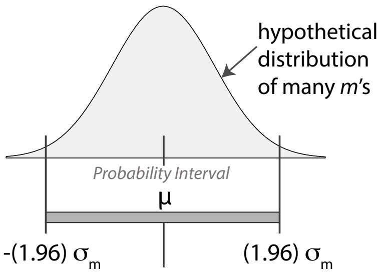


Figure 4.3: Theoretical 95% probability interval that contains 95% of all possible sample means, m, for the variable of interest, calculated for samples each of size n.

The 95% probability interval is 1.96σ on either side of μ. This probability interval becomes the basis for the confidence interval. Both intervals are of the same width but centered over different values.

4.2.4 The Confidence Interval from the Probability Interval

statistical inference: Estimate the value of a population characteristic for a variable of interest.

The calculation of a probability interval presumes that the value of μ is already known, from which the known range of variation of m is then calculated for a specified percentage of its values, such as 95%. Here our primary concern, however, is applying *statistical inference* in the form of a confidence interval to estimate the *unknown* value of the population mean μ.

The logic of the 95% confidence interval follows from the probability interval around the true mean μ that contains 95% of all m's, as illustrated in Figure 4.3. Against the backdrop of what happens *hypothetically* over the means from the repeated samples, get one *actual* sample of data values for the variable Y and then compute m from these data values. Suppose the value of m from the sample obtained shown in Figure 4.4 happens to be a little less than μ.

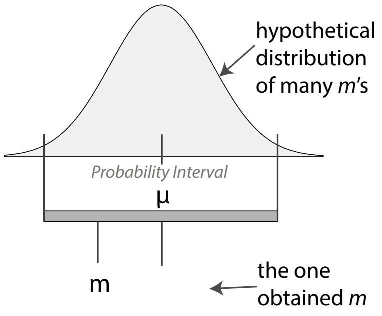


Figure 4.4: One observed value of the sample mean, m, from the one actual sample.

Key Insight: If the interval constructed about μ contains m, then an interval of the *same width* about m contains μ.

The confidence interval is the transfer of the probability interval centered over μ to an interval of the same width centered over m. The key insight is that if the interval constructed about μ contains m, then an interval of the *same width* about m contains μ, as shown in Figure 4.5.

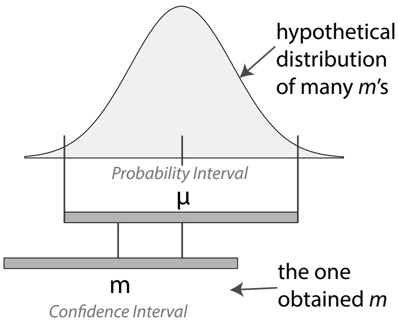


Figure 4.5: The 95% confidence interval for this sample derived from the 95% probability interval.

The interval about m in Figure 4.5 is the confidence interval, in practice constructed *without* knowledge of μ. The analysis of this one sample of data provides the value of the sample mean, m, and the estimated range of sampling variation of m over many hypothetical samples based on the standard error of m. The application of the Central Limit Theorem ensures that m is approximately normally distributed, except in small samples from skewed populations.

Appendix: CLT, Section 4.6, p. 178

Figure 4.6 provides the perspective of the data analyst, who does not know the value of μ. The analyst does not know if the one observed m is larger than or smaller than the value of the unknown μ, nor how much larger or smaller.

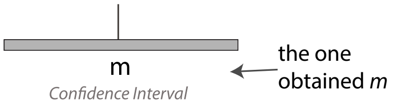


Figure 4.6: The 95% confidence interval 1.96σ on either side of m.

The expression of the 95% confidence interval in terms of the population standard error follows.

$$\text{Theoretical 95\% Confidence Interval:} \quad m \pm (1.96)(\sigma_m), \quad \text{where } \; \sigma_m = \frac{\sigma}{\sqrt{n}}$$

The problem of applying this expression for the confidence interval in practice is that if the population mean, μ, is not known, then neither is the standard deviation, σ. Without σ, the population standard error of the mean, σ_m, cannot be calculated. That is why this calculation is labeled "theoretical". How to compute the confidence interval in practice is explained next.

4.3 Construct the Confidence Interval

4.3.1 The t-distribution and Estimated Standard Error

To infer the value of μ from data, modify the theoretical expression for the standard error of the mean. Replace the actual but unknown standard deviation, σ, with its estimate from a *single* sample from that population, the standard deviation of the sample data, s.

Estimated (sample) standard error of m: $s_m = \dfrac{s}{\sqrt{n}}$

Replacing σ with s in the expression for the standard error results in the *estimated* standard error, s_m, used in place of the true but unknown actual standard error, σ_m.

t-value: Number of *estimated* standard errors that separate a value of m from μ.

The estimation of σ_m with s_m introduces a new statistic in the analysis of the variation of the sample mean, m, expressed in terms of standardized values, applicable to any unit of measurement. When using s in place of σ, instead of a z-value we have a t-value, the number of *estimated* standard errors that separate a value of m from μ.

t-value of the sample mean: $t_m = \dfrac{m - \mu}{s_m}$

A t-value is a standardized value that closely resembles a z-value, but the substitution of s for σ, and thus s_m for σ_m, does change its distribution.

The 95% probability interval for z_m ranges from $-z_{.025} = -1.96$ to $z_{0.25} = 1.96$, but what about the corresponding interval for t_m? To establish the probability interval for a distribution of values of t_m, consider many, many (usually hypothetical) samples all of the same size n from the same population with known mean μ. For each sample calculate both m and s, where m is presumed normally distributed, and then t_m.

To what extent does a distribution of t-values correspond to a normal distribution? The distribution of t_m is also a bell-shaped curve, but the additional source of variability in each calculation implies that a distribution of t_m is wider, more variable than the normal distribution of z_m, which uses the fixed population value σ in place of the corresponding data estimate s.

The distinction between the two families of distributions is the price to pay for needing two estimates from a sample to calculate t_m, m and s, instead of only the one estimate for z_m, m. Without knowledge of the population standard deviation, σ, the corresponding confidence interval is wider than if its value was known. In applications, however, if the population mean, μ, is not known, then neither is the population standard deviation, σ.

Denote the cutoff values for the resulting distribution of t_m, such as for the range of variation for the 95% probability interval, as $-t_{.025}$ and $t_{.025}$. The 95% range of variation of t_m is larger than the corresponding range of variation with normal curve probabilities, so $t_{.025} > z_{.025} = 1.96$. How much larger are the t-distribution cutoffs, such as $t_{.025}$ than the corresponding standardized normal curve values, such as $z_{.025}$?

The mathematicians have done the hard work by providing formulas for the exact cutoffs needed to define the range of variation of the t-statistic, which are provided by computer applications such as `R` and Excel. Another consideration is that there is a different distribution of t-values for each sample size n. The estimate of σ tends to improve as n, or more specifically, the *degrees of freedom*, $n - 1$, for the calculation of s, increases. So as $n - 1$ increases, the variability of the estimate s decreases, yielding a distribution of t_m closer to the distribution of the normal distribution z_m.

degrees of freedom: for each t-distribution, $df = n - 1$.

Each t-distribution yields its own specific cutoff values. The `lessR` function `prob_tcut()` illustrates the relationship between the t and z distributions for any normally distributed variable.

> `lessR` function `prob_tcut()`: t-distribution probabilities.
>
> *Required*: `df`, degrees of freedom, $n - 1$
>
> *Default*: `alpha=0.05` for a 95% probability interval

A sample of size $n = 11$ implies a degrees of freedom of $df = n - 1 = 10$. Figure 4.7 shows a t-distribution slight wider than the normal distribution of z values, with a corresponding cutoff of $t_{.025} = 2.228$.

```
prob_tcut(df=10)
```

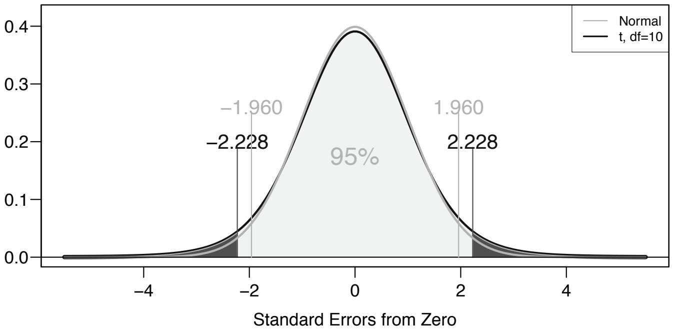


Figure 4.7: t-distribution, $df = 10$, $t_{.025} = 2.228$, and 95% probability interval.

For large sample sizes, the corresponding t-cutoff for the 95% range of sampling variability becomes close to the normal cutoff of 1.96. Consider $n = 201$, that is, $df = n - 1 = 200$, which yields $t_{.025} = 1.972$ according to Figure 4.8. The t-distribution curve almost overlaps the corresponding z-distribution curve.

```
prob_tcut(df=200)
```

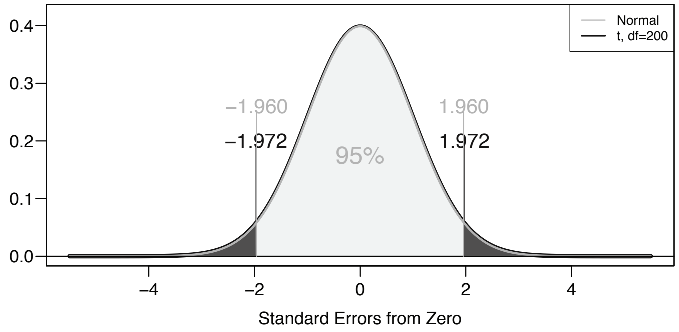


Figure 4.8: t-distribution, $df = 200$, $t_{.025} = 1.972$, and 95% probability interval.

Table 4.2 summarizes the relation of t-cutoffs to the normal cutoff for the range of sampling variation defined by the 95% probability interval.

df	3	5	10	15	20	30	60	100	200	1000	normal
$t_{.025}$	3.182	2.570	2.228	2.131	2.086	2.042	2.000	1.984	1.972	1.962	1.960

Table 4.2: t-cutoffs for the 95% range of variation, which illustrate the penalty compared to the 1.96 reference value when computing the confidence interval from s instead of σ.

$t_{.025}$: About 2, a little higher for $n < 60$, and no lower than $z_{.025} = 1.96$.

For the 95% range of variation, $t_{.025} > z_{.025} = 1.96$, but as Table 4.2 shows, except for very small sample sizes, a good approximation is $t_{.025} \approx 2$. The 95% probability interval based on $z_{0.025}$ provides the actual range of variation of m. Our *knowledge* of this range, however, is based on the larger $t_{0.025}$, which yields larger probability intervals.

4.3.2 Computation

process mean, Section 2.3, p. 47

stable process, Section 2.3, p. 47

run chart, Section 2.3.1, p. 47

control chart, Section 3.3.1, p. 68

Only analysis of a *process mean* provides useful results to the decision maker. Calculate the process mean from data values all generated by the same, unitary process, what is called a *stable process*. If possible, if the data are ordered by the time of their collection, evaluate the stability of the process with a *run chart* or *control chart* before estimating the mean of that process. As previously discussed, the data values from a stable process for a variable exhibit random variation about the mean absent any delineated structure or trend.

The confidence interval is the probability interval centered over the sample mean, m. Because σ is unknown, construct the confidence interval from the t-distribution. The construction of the confidence interval proceeds from the sample summary statistics: n, m and s.

The confidence interval is the sample mean ± the margin of error. The t-statistic indicates how much the standardized mean calculated from the estimated standard error varies. The margin of error expresses our knowledge of how much the corresponding m varies (realizing that m actually varies less, based on the population value of σ_m, but this value cannot be directly accessed).

The margin of error, E, for a 95% confidence interval, is $t_{.025} \approx 2$ standard errors. Obtain the relevant t-distribution cutoff to calculate the margin of error for the confidence interval.

Margin of error: $E = (t_{.025})(s_m)$

The lower bound of the interval is $m - E$ and the upper bound is $m + E$.

margin of error, formula: $E = (t_{.025})(s_m)$

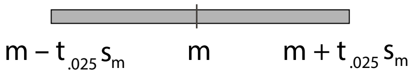


Figure 4.9: The 95% confidence interval.

Example of the confidence interval, Section 4.1.2, p. 78

To illustrate, consider the previous example of the clean up costs after surgery.

Begin with the descriptive statistics.

Data Summary: $n = 100$, $m = \$2984.17$, $s = \$42.59$

From this sample information, and knowledge of the corresponding t-cutoff, compute the confidence interval.

Estimated standard error: $s_m = \dfrac{s}{\sqrt{n}} = \dfrac{42.59}{\sqrt{100}} = \$4.259$

$\mathbf{t_{.025}}$-cutoff. $df = 100 - 1 = 99, \quad t_{.025} = 1.984$

Margin of Error: $E = (t_{.025})(s_m) = 1.984(4.259) = \$8.451$

The confidence interval is the sample mean, m, plus and minus the corresponding margin of error.

Confidence interval:

Lower Bound: $m - E = 2984.17 - 8.451 = \$2975.719$ days

Upper Bound: $m + E = 2984.17 + 8.451 = \$2992.621$ days

The 95% confidence interval of clean up costs ranges from \$2975.72 to \$2992.62.

4.3.3 Meaning

As we have seen, according to the classical (frequentist) model of statistics, the meaning of an inferential analysis follows from the results of many, many, usually hypothetical, samples. The value of the population mean, μ, is some specific value, some constant such as 21 or -4.97. The sample mean varies from sample to sample, so the location of each interval varies. The sample standard deviation varies from sample to sample, so the width of each interval varies. Accordingly, the estimate of μ, the confidence interval, randomly varies from sample to sample. In practice, only one sample and one corresponding confidence interval is observed. Without additional information there is no way to know if that one confidence interval does,

Meaning of the 95% confidence interval: On average, over many samples, each of the same size, 95% of the corresponding confidence intervals contain μ.

or does not, contain μ.

Computer simulation easily allows us to explore the results from multiple samples. To obtain many confidence intervals of the mean from repeated sampling of simulated data, all from the same normal population, use the `lessR` function `simCImean()`. This function is pedagogical. It illustrates the meaning of a statistical concept, and is not used for data analysis per se.

> `lessR` function `simCImean()`: Generate multiple confidence intervals
>
> *Required*: `ns`, number of samples; `n`, size of each sample
> *Default*: `mu=0`, population mean
> `sigma=1`, population standard deviation

The following function call to `simCiMean()` generates 50 samples, each of size $n = 10$ from a normal population with $\mu = 65.5$ and $\sigma = 2.5$. These characteristics describe the distribution of adult USA women's heights.

```
simCImean(ns=50, n=10, mu=65.5, sigma=2.5)
```

The output appears in Figure 4.10.

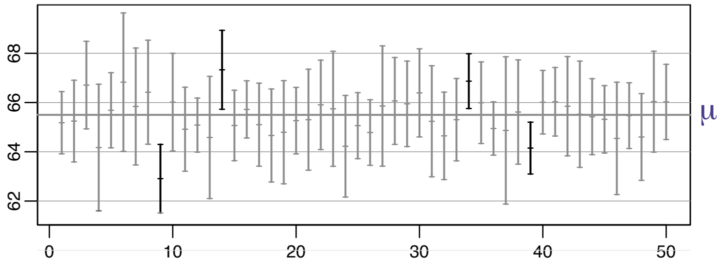


Figure 4.10: 50 random confidence intervals from 50 random samples, $\mu = 65.5$, $\sigma = 2.5$.

For these particular 50 random samples, each of size $n = 10$, 4 of the 95% confidence intervals, or 8%, do not contain μ. For each sample, m, and also s and so s_m, differ, so both the center and width of *each* confidence interval also differ. In actual data analysis, only one of these confidence intervals is observed, without knowledge of μ.

Note that the confidence level of 95% has a different meaning than does the concept of probability, which refers to *future*, random events. Stated another way, the probability that any one random 95% confidence interval to be sampled *will* contain μ is 95%. However, *after* the data have been gathered and the one specific confidence interval has been obtained, then either μ lies in the obtained interval or it does not. After computing the confidence interval, the concept of probability does not apply. Instead, properly expressed results in terms of the 95% confidence level.

An application of the confidence interval follows.

4.4 Application

4.4.1 Background

Problem: The past, assessed by descriptive statistics, is just one arbitrary sample replete with sampling error.

Context. Management has set the criterion that the average supplier ship time should be no more than 7.5 days from the generation of the electronic purchase order. Assess past conformance to this criterion by analyzing past delivery times. Management, however, makes decisions regarding the future. The past can guide the future, but what occurs in the future is of primary importance. Hence the need for statistical inference.

The expression of this criterion for the average ship time is a specification of the value of the underlying population mean: $\mu \leq 7.5$. If the same process continues into the future, then the ship times are random deviations about the process mean. The forecast for future ship times is the estimate of μ, which should be less than 7.5 days.

Response to the problem: Use inferential statistics to estimate μ.

Analytic technique. Estimate the supplier's population mean ship time from last year's shipments with a 95% confidence interval, presuming the underlying assumptions of this procedure are satisfied. If 7.5 is less than all the values of the confidence interval, then the true mean is likely larger than 7.5 and we conclude that the criterion would be satisfied. If 7.5 is greater than all the values of the confidence interval, then the true mean is likely less than 7.5, and we conclude that the criterion is not satisfied. If 7.5 lies within the confidence interval, then it is a plausible value, but then so also would values less than 7.5 and greater than 7.5. If 7.5 is within the interval, the results are equivocal.

Data. The 15 ship times for the past 12 months for a supplier are available on the web, in the file called shiptime.csv. The variable of interest is Time. Read the data into an R data table (data frame) with the lessR function Read().

Read() function, Section 1.2.3, p. 13

```
d <- Read("http://lessRstats.com/data/shiptime.csv")
```

4.4.2 Assumptions

Process stability. The first task in the analysis is to evaluate process stability, if possible. The primary issue here is that all the sampled data should be from the same population with mean μ. Including data from different shipping procedures with different values of μ would diminish the forecasting accuracy of future events from the *current* process.

evaluate process stability, Section 2.3.1, p. 47

To evaluate the stability of the process, examine the data for the 15 sequentially ordered shipments with the run chart with the lessR function Plot(). Set the run parameter to TRUE to indicate a run chart plot.

Plot() function, Section 2.3.1, p. 47

```
Plot(Time, run=TRUE, xlab="Shipment")
```

The line chart appears in Figure 4.11.

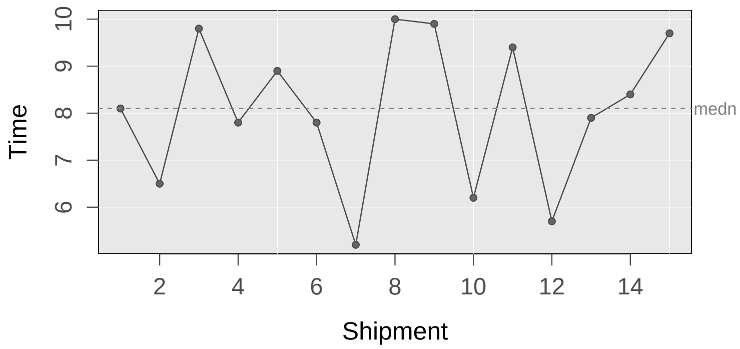


Figure 4.11: Run chart of the 15 ship Time data values.

Over time, the plot shows no pronounced pattern for the 15 ship times, just apparent random variation about a stable mean with about the same level of variation across all the ordered data values.Conclude stability of μ and σ for all sampled values. That is, assume that the same fulfillment process generates all 15 data values, and so becomes suitable for forecasting future performance of that process. Only values generated by the process projected into the future are eligible for inclusion in the analysis of the process mean.

Histogram() function, Section 1.3.2, p. 21

histogram of ship times, Section 1.3.3, p. 24

histogram bin shift, Section 1.3.3, p. 24

Normality. The sample size, $n = 15$, is less than 30, so evaluate the histogram of the data values for normality of the data. According to the central limit theorem, if the data indicate approximate normality, or at least not much skewness, the sample mean, m, is at least approximately normally distributed regardless of the sample size, n. The histogram of ship Time, adjusted for the bin shift artifact by moving the start point from the default of 5.0 to 4.5, is shown in Figure 4.12.

```
Histogram(Time, bin_start=4.5)
```

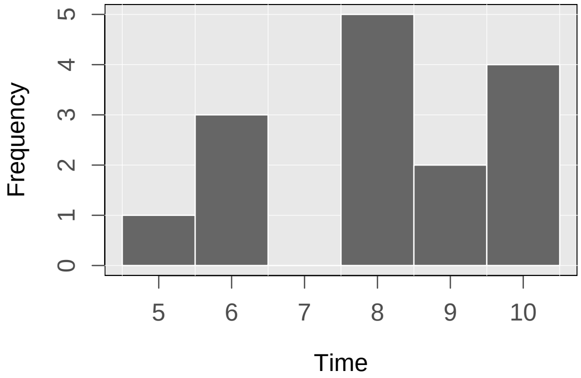


Figure 4.12: Histogram of 15 ship Time values with the bins starting at 4.5.

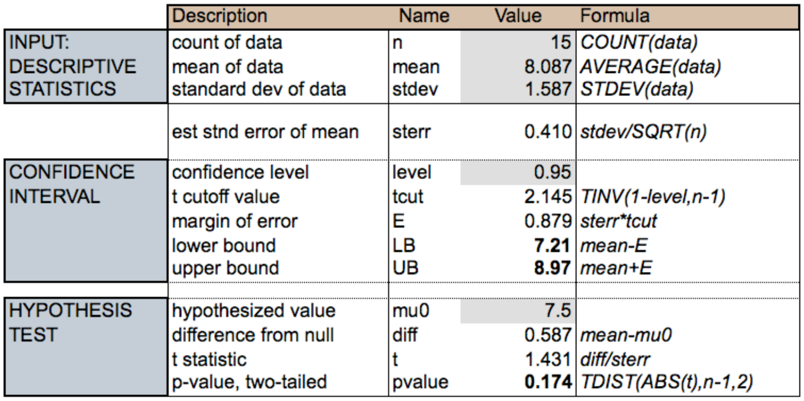
This histogram of the *same* data is “more normal”, and so provides support for a normal corresponding sample mean, m. A sample size of $n = 15$ is small, and most samples of that size from a normal population are not going to exhibit perfect normality.

4.4.3 Computation

Excel. An Excel worksheet for the computations for the confidence interval of the mean is also available for download. (This worksheet also provides for other calculations, the hypothesis test and needed sample size, discussed later.)

`http://lessRstats.com/excel/MeanInference.xls`

One possibility is to compute the descriptive statistics – sample size, mean and standard deviation – in this worksheet. There is a range of values all in one column named `data`. The Excel functions `COUNT()`, `AVERAGE()` and `STDEV()` provide these three summary statistics, from the `data` range of values. Redefine this range for a new set of data, or just enter the cell range in directly in the respective function calls. Or, if already available, just enter the sample values directly into these three cells. Also specify the confidence level. For the given $df = n - 1$, the Excel `TINV()` function for t-inverse provides the corresponding t cutoff values.



	Description	Name	Value	Formula
INPUT: DESCRIPTIVE STATISTICS	count of data	n	15	*COUNT(data)*
	mean of data	mean	8.087	*AVERAGE(data)*
	standard dev of data	stdev	1.587	*STDEV(data)*
	est stnd error of mean	sterr	0.410	*stdev/SQRT(n)*
CONFIDENCE INTERVAL	confidence level	level	0.95	
	t cutoff value	tcut	2.145	*TINV(1-level,n-1)*
	margin of error	E	0.879	*sterr*tcut*
	lower bound	LB	**7.21**	*mean-E*
	upper bound	UB	**8.97**	*mean+E*
HYPOTHESIS TEST	hypothesized value	mu0	7.5	
	difference from null	diff	0.587	*mean-mu0*
	t statistic	t	1.431	*diff/sterr*
	p-value, two-tailed	pvalue	**0.174**	*TDIST(ABS(t),n-1,2)*

R. With `R`, calculate the confidence interval of the mean from data or summary statistics with the `lessR` function `ttest()`, or use the simpler version `tt_brief()`. The output appears in Listing 4.5.

```
ttest(Time)   or   ttest(n=15, m=8.09, s=1.59)
```

4.4.4 Results

Interpretation. The interpretation of the confidence interval is straightforward: With 95% confidence, the true average shipping time for this supplier is somewhere between $7\frac{1}{5}$ and 9 days. Rounding the results provides an interpretation more easily understood without any meaningful sacrifice of precision.

template for confidence interval interpretation, Section 4.1.2, p. 80

Note that the interpretation is not a simple verbal restatement of the lower and upper bounds of the interval: The 95% confidence interval for average ship time is 7.21 to 8.97 days. This statement is *not* an interpretation because it presumes the intended

```
------ Description ------

n = 15,  mean = 8.09,  sd = 1.59

------ Inference ------

t-cutoff: tcut = 2.145
Standard Error of Mean: SE = 0.41

Margin of Error for 95% Confidence Level: 0.88
95% Confidence Interval for Mean: 7.21 to 8.97
```

Listing 4.5: Output of `ttest()`.

audience understands the meaning of the technical phrase "95% confidence interval". An effective *interpretation* conveys the meaning of the results to a non-technical audience, free of all jargon.

Interpretation: Conveys the meaning of results free from technical jargon.

Conclusion. Our conclusion begins with the descriptive statistics, but the primary result is the conclusion of the inferential analysis. In terms of the data summary, average delivery time of last year's 15 deliveries was 8.09 days, yet management will only accept an average of 7.5 days. For the inferential analysis, the confidence interval, an average actual delivery time as low as 7.21 days is plausible. The data analysis does not definitively answer if the goal of an average of 7.5 days has been achieved by that supplier for last year's shipments.

key concept: Follow the interpretation of the results with the conclusion of the analysis and then the managerial decision.

Decision. From this analysis the managerial decision, the purpose of the analysis, follows. Before finding another supplier, provide more opportunity to demonstrate a sufficiently small delivery time, either with more data or analysis and potential improvement of the supplier's internal shipping process.

4.5 Confidence Level and Margin of Error

4.5.1 Confidence Level

The previous examples of confidence intervals had confidence levels of 95%. Yet the 95% level is only one of many possible confidence levels, albeit perhaps the most widely encountered level. Set the confidence level according to the range of variation defined by the t-cutoff value. For a given t-distribution, each t-cutoff value corresponds to a different range of sampling variation, such as perhaps the other relatively common confidence levels of for 90% and 99%. Consider first the corresponding z-cutoff values, which set the baseline for the slightly larger t-cutoffs.

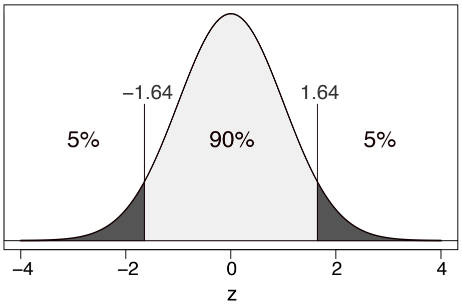


Figure 4.13: Normal Distribution: 90% Range of Variation, within 1.64 standard errors of the population mean, μ

90% of the values of any normal distribution, including for m, fall within 1.64 standard deviations of the mean, μ.

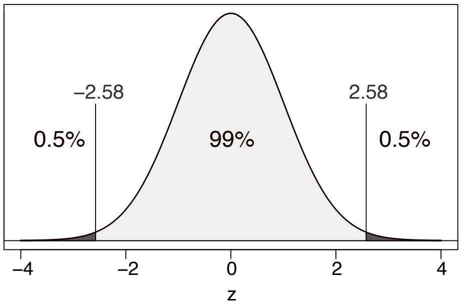


Figure 4.14: Normal Distribution: 99% Range of Variation, within 2.58 standard errors of the population mean, μ.

99% of the values of any normal distribution, such as for a normally distributed m, fall within 2.58 standard deviations of the mean, μ.

Use the corresponding t-cutoff to set the confidence level. Each t-cutoff is a little larger than the corresponding z-cutoff, as shown in Table 4.3.

Level	Interval	t-cutoff
90%	$m \pm t_{.05}(s_m)$	$t_{.05} > 1.65$
95%	$m \pm t_{.025}(s_m)$	$t_{.025} > 1.96$
99%	$m \pm t_{.005}(s_m)$	$t_{.005} > 2.57$

Table 4.3: Relationship of the t-cutoff to the confidence level.

How to choose among the different confidence levels when constructing a confidence interval? The ideal confidence interval achieves two desirable goals.

Ideal Confidence Interval

- High level of confidence of containing μ.
- Narrow margin of error, E.

Unfortunately, for the same data set, satisfy one goal only at the expense of the other. Consider first the pursuit of a high level of confidence.

Purpose	Description
Goal	High confidence the CI contains unknown μ
Action	To increase probability of containing the unknown value of μ, move the confidence level from 95% to 99%
Result	Larger confidence level results in wider interval, and so a larger margin of error

Table 4.4: Relationship of the t-cutoff to the confidence level.

Or, pursue the other desirable characteristic of a confidence interval.

Purpose	Description
Goal	Low margin of error, i.e., precision
Action	To decrease size of confidence interval, decrease confidence level from 95% to 90%
Result	Narrower interval yields a lower probability that the interval actually contains μ

Table 4.5: Relationship of the t-cutoff to the confidence level.

Recalculating a confidence interval to obtain either more confidence or a narrower interval results in a *less desirable* value of the other characteristic. This trade off between competing goals implies that the chosen confidence level is to some extent arbitrary. Each choice of confidence level provides the *same* amount of information, balancing the margin of error against the confidence level. So usually choose 95%, which is both a high level of confidence and a "nice" number, divisible by 5, that still gets most of the values.

ttest() function, Section 4.4.3, p. 95

The lessR function ttest() or its simpler version tt_brief() calculates the confidence interval with a default confidence level of 95%. To explicitly specify the confidence level, use the conf_level parameter, such as with the following function calls.

```
ttest(Y, conf_level=.90)
ttest(Y, conf_level=.99)
```

Here the different confidence intervals are calculated for the variable Y. Note that the confidence level is specified with a proportion, such as 0.90, and not a percentage, such as 90.

Table 4.6 organizes the output from these analyses, plus the original analysis with the 95% level of confidence. Consider the previous application that analyzed the 95% confidence interval for the average ship time of a supplier. The data summary is $n = 15$, $m = 8.087$ days, $s = 1.587$ days.

The narrowest width, of 1.44 days, comes with the least confidence, .90. Moving the confidence level up to .99 increases the confidence interval's width all the way to 2.44 days.

Level	$t_{.025}$	LB	UB	Width	
.90	1.761	7.36	8.81	1.44	
.95	2.145	7.21	8.97	1.76	
.99	2.977	6.87	9.31	2.44	

Table 4.6: Width of the confidence interval and confidence level.

4.5.2 Choose the Needed Sample Size

Distinguish between the extent of the margin of error that you obtained and what you want. What you have is E, the obtained margin of error *obtained* with the sample of size of n. What you want is $E_{desired}$, the margin of error *desired*, which may require an increase of the current sample size, n, to the larger value n_{needed}.

$E_{desired}$: Desired margin of error.

How large a sample is needed to obtain the smaller margin of error that is desired? As always in life, including statistics, there are no guarantees. The goal is to have a very high probability that when the new confidence interval is calculated from the larger sample, the desired margin of error will be obtained. Calculate the needed larger sample size, n_{needed}, so that the new 95% confidence interval has a .90 probability of obtaining the desired margin of error, $E_{desired}$.

The procedure is to move from the initial sample to the larger sample to obtain a diminished margin of error.

1. Specify the desired margin of error (precision), $E_{desired}$.
2. Obtain initial data sample, n.
3. Calculate margin of error, E, then the confidence interval.
4. If $E_{desired} < E$, calculate n_{needed}.
5. Gather new data for larger sample.
6. Re-calculate the margin of error, E, and the confidence interval.

The end result is a larger sample with a lower margin of error.

Calculate the needed sample size with a two-step procedure. First calculate the preliminary estimate of sample size, n_s from the initial sample. With the sample standard deviation, s, apply the following expression.

$$n_s = \left[\frac{(1.96)(s)}{E_{desired}}\right]^2$$

Second, calculate the actual estimate of sample size, n_{needed}. The issue here is that the sample standard deviation, s, may underestimate σ, so revise n_s upward for a given probability[1] of obtaining $E_{desired}$ with a 95% level of confidence.

.70 probability: $n_{needed} = 1.054n_s + 4.532$
.90 probability: $n_{needed} = 1.132n_s + 7.368$ ▷ Most often used
.99 probability: $n_{needed} = 1.242n_s + 10.889$

[1] These coefficients are derived from analysis of a paper by Kupper and Hafner in the *American Statistician*, 43(2):101-105, 1989.

The adjustment explicitly accounts for the variability of the standard deviation estimate, s, across repeated samples. In practice, the standard error of the sample standard deviation is quite large in small samples. So by chance the one obtained sample standard deviation, s, may be a very large underestimate of the true population standard deviation, σ. The result would be a value of estimated sample size that would be too small to obtain the desired margin of error with any reasonable probability. Without this adjustment the actual probability of getting the desired margin of error is quite a bit less than .90.

ttest function, Section 4.4.3, p. 95

To specify the needed sample size invoke the parameter `Edesired` for the `lessR` function that calculates the confidence interval of the mean, `ttest()`. Here consider a value of $E_{desired} = 0.5$.

```
ttest(Time, Edesired=0.5)
```

This example provides the 95% confidence interval for the variable Time as well as the needed sample size for a desired margin of error of 0.5.

4.5.3 Application

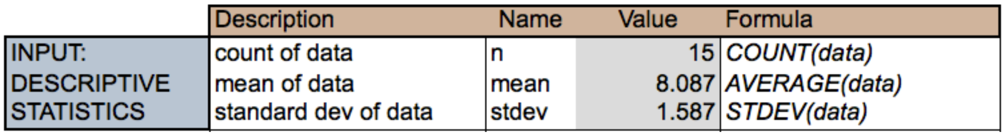
Confidence interval for ship time, Section 4.4, p. 93

Reconsidering Ship Times. The previous application of a confidence interval CI of the ship time of a supplier estimated the population mean ship time to be between 7.21 to 8.97 days with 95% confidence. Unfortunately, this interval is deemed too large, achieving an obtained margin of error, E, of only 0.879 days. This result is equivocal regarding the criterion that the true mean ship time be less than 7.5 days. Perhaps the criterion was not reached such that the true mean is larger than 7.5 days, yet 7.5 is within the interval and so is plausible.

This confidence interval for ship time involved only the 15 last shipments from that company. Unfortunately, management deemed the precision of estimation as not sufficiently precise. Instead, management specifies a maximum desired margin of error of only one-half day. How many shipments need be sampled to reach a .90 probability of obtaining a margin of error of half a day or less at the 95% confidence level?

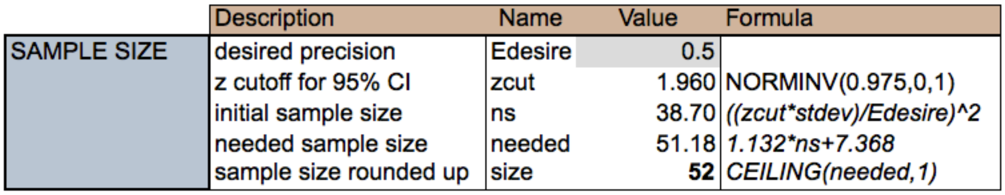
Excel worksheet, Section 4.4.3, p. 95

Excel. The previously introduced Excel worksheet for calculating a confidence interval also contains the calculations for the needed sample size. Begin with the summary statistics as previously entered. From the initial sample of $n = 15$ shipments, $s = 1.587$, with $E_{desired}$ specified as 0.5.



	Description	Name	Value	Formula
INPUT: DESCRIPTIVE STATISTICS	count of data	n	15	*COUNT(data)*
	mean of data	mean	8.087	*AVERAGE(data)*
	standard dev of data	stdev	1.587	*STDEV(data)*

The task is to obtain the 0.9 probability of getting desired margin of error of 0.5 days.



	Description	Name	Value	Formula
SAMPLE SIZE	desired precision	Edesire	0.5	
	z cutoff for 95% CI	zcut	1.960	NORMINV(0.975,0,1)
	initial sample size	ns	38.70	*((zcut*stdev)/Edesire)^2*
	needed sample size	needed	51.18	*1.132*ns+7.368*
	sample size rounded up	size	**52**	*CEILING(needed,1)*

The calculated needed sample size is 52.

R. Specify the value of $E_{desired}$ the option `Edesired` for the `lessR` function `ttest()`.

```
ttest(Time, Edesired=0.5)
```

The output for the needed sample size appears in Listing 4.6.

```
Desired Margin of Error:  0.50

For the following sample size there is a 0.9 probability of obtaining
the desired margin of error for the resulting 95% confidence interval.
-------
Needed sample size:  52

Additional data values needed:  37
```

Listing 4.6: Needed sample size analysis.

Calculations. Step 1: Initial sample size.

$$n_s = \left[\frac{(z_{.025})(s)}{E_{desired}}\right]^2 = \left[\frac{(1.96)(1.587)}{0.5}\right]^2 = 38.69$$

Step 2: Upward adjust the initial estimate to obtain the actual estimated sample size.

$$\begin{aligned} n_{needed} &= 1.132 n_s + 7.368 \\ &= 1.132(38.69) + 7.368 \\ &= 51.17. \end{aligned}$$

Now round up, 51.17 to $n_{needed} = 52$. From the initial sample of 15, $52 - 15 = 37$ additional values need to be obtained to achieve a sample size of 52.

Conclusion. Collect information on 52 shipments instead of 15, so 37 more shipments are needed. There is a .90 probability that when the revised 95% confidence interval is calculated over all 52 shipments, the resulting margin of error will be 0.5 or less.

The problem with this result is that 52 shipments are not available. And, even if data from several years past were available, the longer the time past since the data were collected, the more likely that the underlying process has changed. Data collected from a different shipping process are invalid for estimating the *current* population mean shipping time, μ, the estimate projected into the future as a forecast subsequent ship times. This analysis simply is not definitive.

4.6 Summary

A set of population values, such as the population mean, μ characterizes various aspects of any process. These population values cannot be directly observed due to the obfuscation of the accompanying sampling error when a sample of data values are obtained to measure these characteristics. To estimate the unobservable value of μ, construct and interpret a confidence interval. This estimate of μ is given with a specified margin of error.

The estimate of μ begins with the sample mean, m. The problem is that the presence of this inevitable sampling error implies that each sample from the same population yields a different sample mean, m. As such, m itself is a variable, and its distribution over usually hypothetical multiple samples is the key information needed to estimate μ. The extent of the fluctuation of the m over many samples is the basis for the confidence interval.

Describe the fluctuation of m across samples by its standard deviation, called the standard error of the mean, given by a simple formula that applies to any distribution.

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of the population from which the sample was obtained is σ, and the standard deviation of all possible values of m is σ_m.

To describe the distribution of a variable, here m, describe its shape and mean in addition to its standard deviation. The Central Limit Theorem shows that, except for very small samples from nonnormal populations, the distribution of m is approximately normal. Further, the population mean of the variable m equals the population mean of the original variable Y, μ. Accordingly, the confidence interval that likely contains μ is based on normal curve probabilities, such as that 95% of all the values of a normal distribution are within 1.96 standard deviations of its mean.

In practice both the population values μ and σ are usually not known, so estimate the standard error of the mean with the sample standard deviation, s.

$$s_m = \frac{s}{\sqrt{n}}$$

Using an estimated standard error to estimate an unknown μ introduces another source of error: both m and s vary from sample to sample. When using s_m, the family of t-distributions provides the cutoff values that define a given range of variation of m, such as $t_{.025}$ for the 95% range of sampling variation. There is a separate t-distribution for each degree of freedom, where, $df = n - 1$.

Construct the 95% confidence interval around the sample mean by moving $t_{.025} \approx 2$ estimated standard errors on either side of the sample mean, m. The true population mean, μ, is contained within this interval at a level of 95% confidence. That is, 95% of all confidence intervals constructed about all samples contain μ. The interpretation of the 95%confidence interval follows this template: With 95% confidence, the true average _____ is between *lb* and *ub*. Here *lb* is the lower bound of the confidence interval and *ub* is the upper bound. Calculate the sample size needed to achieve a desired margin of error for the estimate of μ so that the width of the confidence interval is small enough to be useful.

Concepts

central limit theorem
confidence interval
confidence level
cutoff value
estimated standard error of the sample mean
maximum error
sampling distribution of the mean
sampling error
standard error of the mean
t-value

`lessR` Instructions

```
ttest(Y) or tt(Y): confidence interval for variable Y
tt_brief(Y): briefer analysis for variable Y
```

Appendices

Appendix: Distribution of the Sample Mean

Show the Shape of the Distribution of the Sample Mean. Use computer simulation to illustrate the distribution of m.

`lessR` function `simCLT`

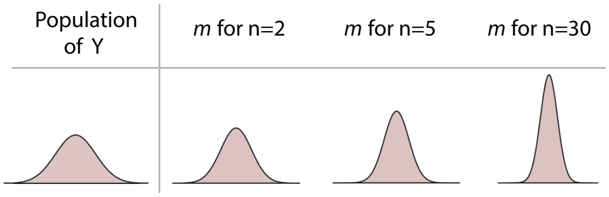
To investigate if m is normal for a given population and sample size, simulate its specified distribution.

Required: `ns`, number of samples
`n`, size of each sample
Required: `dist`: `"normal"`, `"uniform"`, `"antinormal"`, or `"lognormal"`

For example, to calculate m from each of 1000 samples, each with two data values, from a uniform distribution:

```
> simCLT(ns=1000, n=2, dist="uniform")
```

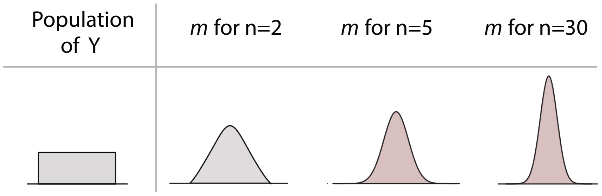
Central Limit Theorem: Normal Data. The normal population, from which the data are sampled. Take many, many different samples, each sample of the smallest possible size, $n = 2$. The many, many means of each of these samples for m is also normal If the sample size is m is also normal. $n = 30$ and larger



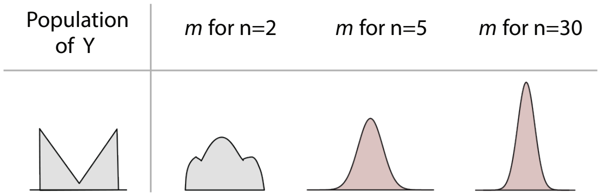
Example of All Possible Sample Means of Size 2. Uniform distribution, 5 equally probable values: 0, 1, 2, 3, 4. There are only 25 possible samples of size 2. There are more ways to get $m = 2$ than $m = 0$ or $m = 4$, so values of m tend to converge toward $\mu = 2$.

Sum	Mean	Possible Samples	Count	Prob
0	0.0	0,0	1	1/25=0.04
1	0.5	0,1 1,0	2	2/25=0.08
2	1.0	0,2 1,1 2,0	3	3/25=0.12
3	1.5	0,3 1,2 2,1 3,0	4	4/25=0.16
4	2.0	0,4 1,3 2,2 3,1 4,0	5	5/25=0.20
5	2.5	1,4 2,3 3,2 4,1	4	4/25=0.16
6	3.0	2,4 3,3 4,2	3	3/25=0.12
7	3.5	3,4, 4,3	2	2/25=0.08
8	4.0	4,4	1	1/25=0.04
Total			25	1.00

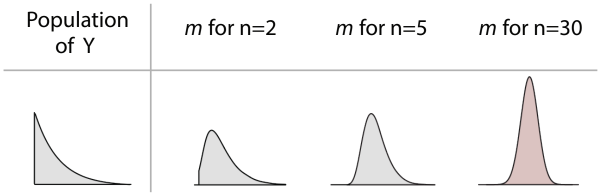
Central Limit Theorem: Uniform Data. The small, medium and large values of Y *equally* likely Take many, many different samples of Y, each sample of the smallest possible size, $n = 2$ The distribution of the sample means for $n = 2$ is *almost* normal If the sample size is m is approximately normal $n = 30$ and larger.



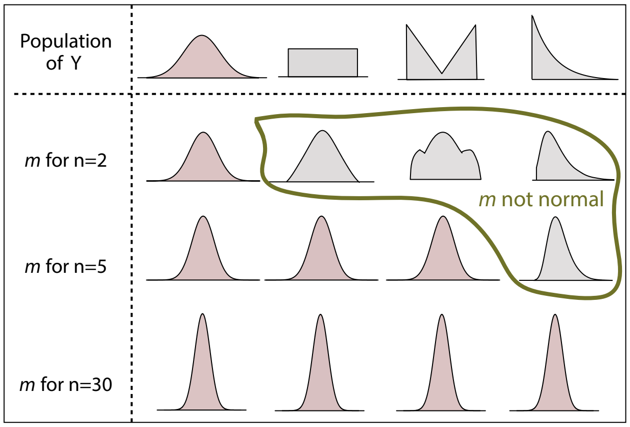
Central Limit Theorem: "Anti-Normal" Data. The population values in the middle *less* likely than tail values Take many, many different samples of Y, each sample of the smallest possible size, $n = 2$ The distribution of the sample means for $n = 2$ is almost normal If the sample size is m is approximately normal $n = 30$ and larger.



Central Limit Theorem: Skewed Data. A skewed population Y from which the data are sampled. Take many, many different samples of Y, each sample of the smallest possible size, $n = 2$ The distribution of the many, many means of each of these samples is also skewed If m's still retains some skew, then $n = 30$ and larger.



Central Limit Theorem: Summary. The sample mean m is $\ldots n$ from 2 onward n less than 5 n is at least 30



CLT: Conclusion and Practical Consequences. The sample mean, m, must be at least approximately normal to apply normal (or related) distribution probabilities for the computation of the confidence interval. IF sample size $n > 30$, then the sample mean, m, is at least approximately normally distributed *unless* the data are sampled from a severely skewed population.

Particularly for smaller sample sizes, the histogram or other frequency display should be checked to evaluate the skewness of the population distribution. If the population is somewhat or moderately skewed, then a sample size of at least $n = 30$ should be obtained. For severely skewed populations, the use of normal curve probabilities on which to base statistical inference may not be acceptable.

For symmetric populations, the sample size may be as small as $n = 10$ or even $n = 5$ to properly employ normal curve probabilities. If the population is normal, then the resulting distribution of m is normal even for the smallest possible sample size of $n = 2$. A small sample size will generally result in poor estimation, but in these circumstances, a normal or at leasts symmetric distribution, results in the valid application of normal curve probabilities.

Knowing that m usually follows a normal distribution is a powerful and interesting result that leads to a world in which many process, which themselves are created from the sums of constituent components, are normally distributed. But how do we use this information for statistical inference?