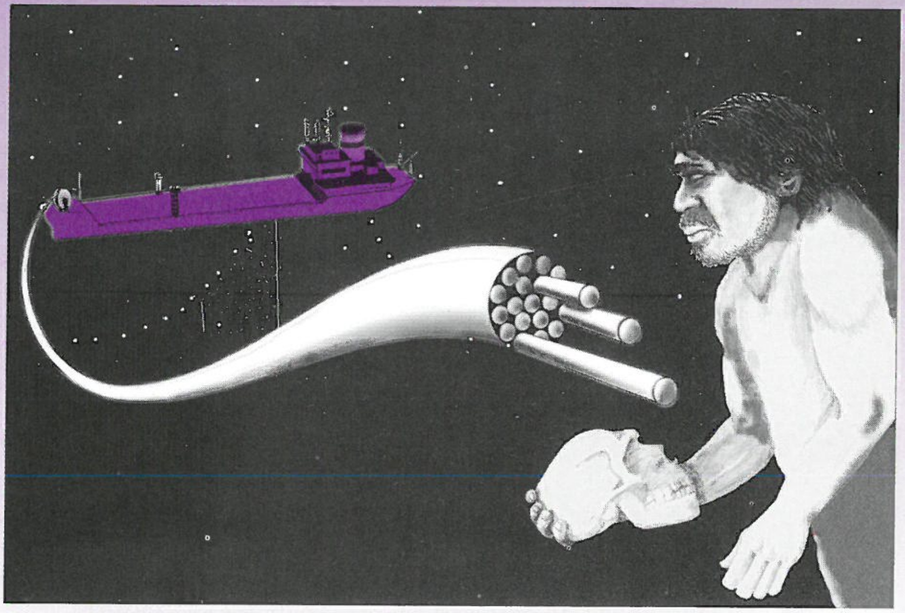
CHAPTER 7

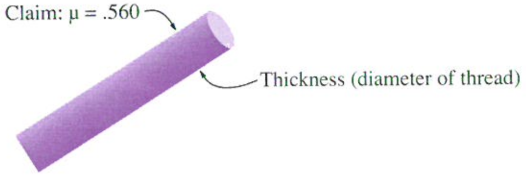
Hypothesis Testing



7.1 Two-Tailed Hypothesis Tests (Large Sample, $n \geq 30$)

Hypothesis testing is nothing more than a formalized approach to the central limit theorem incorporating the concepts of accept/reject decision making and Type I error. Let's see how it works in the following problem. ▼

Suppose the Fiche Company (a manufacturer of telephone cable) receives shipments of fiber optic thread, hair-thin strands of glass capable of transmitting hundreds of thousands of times more information than a copper wire. The Fiche Company will ultimately coat the fiber-optic threads with steel and plastic and bind several into cables to be laid on ocean floors for intercontinental communications. However, it is important for production purposes that the incoming shipments of hair-thin glass fiber thread maintain an average thickness of .560 mm. Of course the supplier of the thread claims this is so.



This is a typical situation in business. A supplier ships you goods and makes a claim with the expectation that you will believe that claim. In this case, the claim is: the average thickness of fiber optic thread in the shipment is .560 mm. In statistical terms, we call this a hypothesis.

A **hypothesis,** then, is merely a claim put forth by someone. This hypothesis or claim is denoted by the symbol H, or H_0 (H-sub-zero) and referred to formally as the **null-hypothesis.***

In this case, our claim or null hypothesis would be written

$$H_0: \mu = .560 \text{ mm}$$

This null hypothesis may or may not be true. The supplier may have documented evidence for making such a claim, or may simply be guessing. In fact, for all we know, the supplier may be lying outright, which of course obliges us as prudent individuals to test their claim. This test is referred to as a hypothesis test.

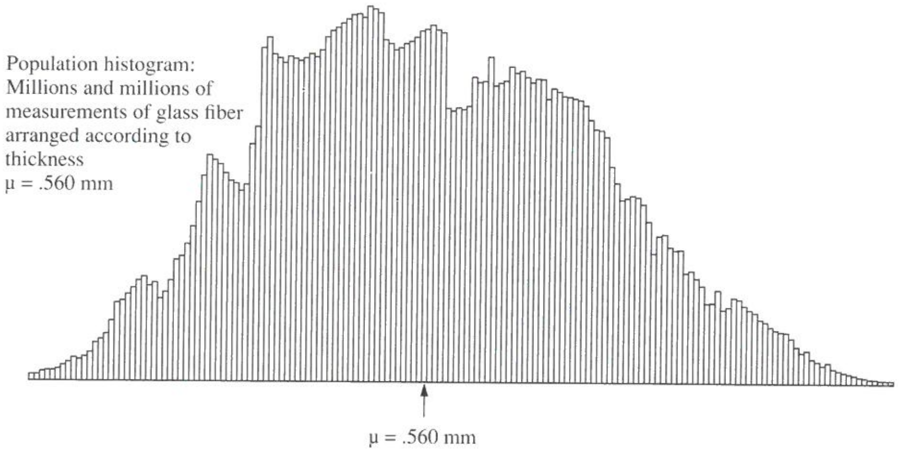
*Technical note: Actually the symbol H_0 originates from tests involving the comparison of two population means or proportions, however the symbol H_0 has now evolved to represent any hypothesis set up for the purposes of seeing if it can be rejected.

Hypothesis Test
A test designed to prove or disprove some initial claim, your null hypothesis, H_0.

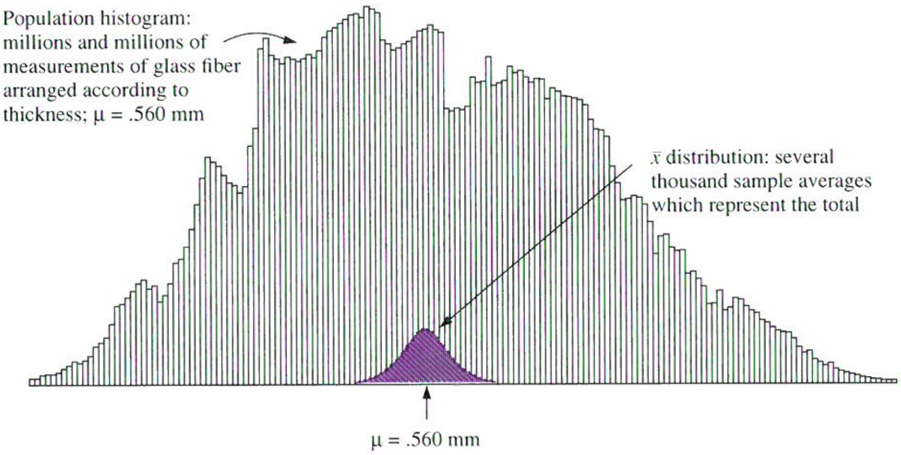
When dealing with a hypothesis test, we always begin by assuming the claim or null hypothesis (H_0) is true, in this case that the supplier is correct, that indeed the average thickness is $\mu = .560$ mm for these shipments of fiber optic thread.

We begin a hypothesis test by assuming H_0 is true.

Indeed, if we accept H_0: $\mu = .560$ mm as true (which we must to begin a hypothesis test), then we know from decades of experience a certain logic will necessarily follow, namely, if we were to measure the thickness of all the glass fiber in the shipment and arrange these measurements according to size into a histogram, these measurements would probably cluster about the average value of $\mu = .560$ mm, however many measurements would be less than .560 mm and many would be more, and the histogram *might* take on the following shape.



Notice this population is somewhat ragged in shape with a slight skew. Although in real life we may not actually know the shape of the population prior to sampling, it would not be unusual for such a ragged skewed shape to appear. Although the output from one process or machine, properly operating and running uninterrupted, is often found to be normally or nearly normally distributed, an entire shipment may very well consist of output from several machines or processes over several periods of time and, thus, could vary considerably. When the output from various processes are mixed, a normal distribution may or may not form, depending on a number of factors. However, this should not make a difference in our analysis of μ, since whatever the shape of your population, as long as the sample size exceeds 30, the $\bar{x}$ distribution will be normally distributed, as follows:



However, we do have another problem.

Noticeably absent from the above histogram is information concerning the standard deviation of this population, σ, which in real-life situations is often not supplied. In fact, more often than not, it is simply unknown. However, without σ we cannot calculate $\sigma_{\bar{x}}$.

$$\text{Remember: } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

And without $\sigma_{\bar{x}}$, we cannot estimate the spread of our $\bar{x}$ distribution, which tells us where we should expect sample averages ($\bar{x}$'s) to cluster—which of course forms the entire basis of our central limit theorem analysis. In other words, we are stuck!

But wait, the problem is not insurmountable. We have learned from prior exercises that when we randomly select 30 or more measurements from a population that

$\bar{x} \approx \mu$ the sample average, $\bar{x}$, is approximately equal to the population average, μ, and

$s \approx \sigma$ the sample standard deviation, s, is approximately equal to the population standard deviation.

If indeed $s \approx \sigma$, that is, the individual measurements in one sample are spread out in a manner similar to how the measurements in the entire population are spread out, we may be able to use the standard deviation of one sample, s, as an estimator of the standard deviation of the entire population, σ. Experience has confirmed that when your sample size is over 30, indeed the spread of measurements in one sample is a good estimator of the spread of measurements in the entire population—that is, s is a good estimator of σ, and this is precisely what is done in industry and research studies.

s is used to estimate σ.

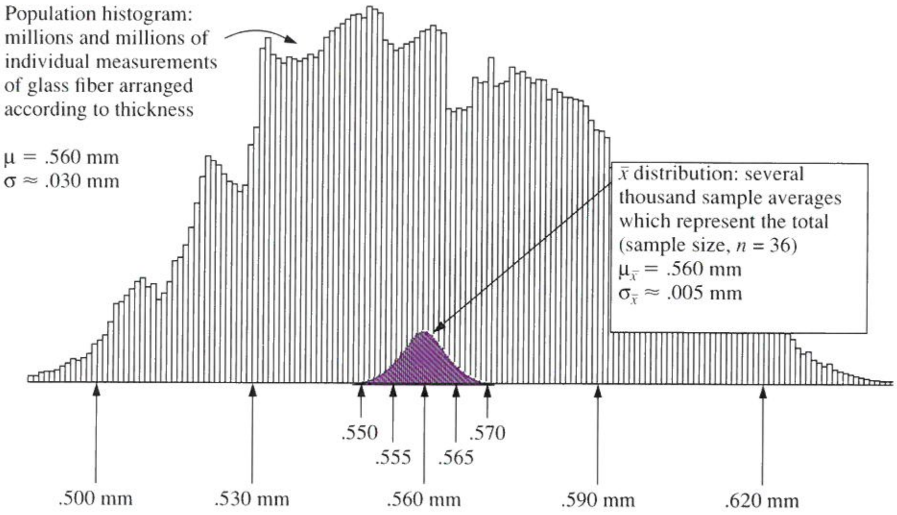
Since the standard deviation of one sample should give us what we want to know, namely, an approximation of σ, the standard deviation of the population, then the telephone cable manufacturer is obliged on receiving the shipment to take a *random sample*. Although many results are possible, let us say, for the purposes of this example that the manufacturer randomly samples 36 pieces of fiber-optic thread and calculates the following:

$$\begin{aligned} n &= 36 \text{ measurements} \\ \bar{x} &= .553 \text{ mm} \\ s &= .030 \text{ mm} \end{aligned}$$

If this is indeed a properly conducted random sample, the *spread* (standard deviation) of the 36 measurements should be similar to the spread (standard deviation) of the entire population. That is, if $s = .030$ mm (note sample results above) and if $s \approx \sigma$, then σ must be approximately equal to .030 mm. And we can use this estimate to calculate $\sigma_{\bar{x}}$, as follows:

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} \approx \frac{.030}{\sqrt{36}} \approx \frac{.030}{6} \\ &\approx .005 \text{ mm} \end{aligned}$$

Now that we know $\sigma_{\bar{x}}$ is approximately equal to .005 mm, we can now estimate the spread of the $\bar{x}$ distribution.



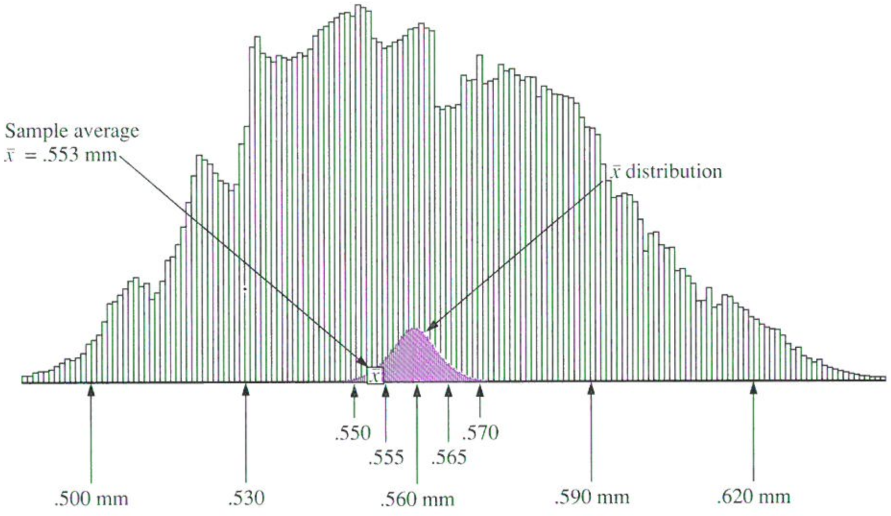
Keep in mind, what we have done so far is a make-believe construction based solely on the assumption that the supplier's claim $\mu = .560$ mm is true. We really do not know whether $\mu = .560$ mm is true or not. We are merely saying: “if” $\mu = .560$ mm is true, and “if” we were to measure every piece of fiber in the shipment, and “if” we continually took random samples of 36 measurements and calculated the sample average, $\bar{x}$, for each sample, then the central limit theorem tells us that the $\bar{x}$'s should form into a normally distributed $\bar{x}$ distribution, symmetrical about $\mu = .560$ mm and spread out as shown above.

Okay, now that we know what the $\bar{x}$ distribution should look like if the supplier's claim is true, how do we prove (or disprove) $\mu = .560$ mm? Simple. We take a random sample of 36 measurements from our shipment, calculate the sample average, $\bar{x}$, and observe if this $\bar{x}$ reasonably fits into the expected $\bar{x}$ distribution.

Wait a minute. We a!ready took a random sample of 36 measurements. True. There's no point spending time and money on another sample. Let's use the $\bar{x}$ we observed from the earlier sample. If you recall, our sample results were as follows (reprinted here for convenience):

$n = 36$ measurements
$\bar{x} = .553$ mm ← (Now we are interested in this measurement)
$s = .030$ mm

Notice that, now, we are concerned with the $\bar{x}$ of the sample. In other words, does this $\bar{x}$ of .553 mm reasonably fit into our expected $\bar{x}$ distribution? And the answer is, yes. We can look at this sample average of .553 mm and look at the $\bar{x}$ distribution and see that this $\bar{x}$ of .553 mm is a reasonably likely occurrence. Observe:



Since an $\bar{x}$ of .553 mm would be a reasonably likely occurrence, we conclude that the supplier's claim ($\mu = .560$ mm) is quite possible. If we choose to make a firm *accept* H_0 or *reject* H_0 decision, then we

$$\text{Accept } H_0\text{: } \mu = .560 \text{ mm}$$

In reality, there is not enough evidence to prove μ is *precisely* .560 mm. The best we can show is that $\mu = .560$ mm is reasonably possible given the evidence of this one sample. The concept of hypothesis testing is much like a jury trial: $\mu = .560$ mm is innocent (accepted) unless *proven guilty*. Since a sample average of $\bar{x} = .553$ mm does not prove the supplier's claim false, then we must assume the supplier's claim is true.

Professionally, this conclusion is written in a number of ways. Two of the most popular are:

The null hypothesis cannot be rejected

or

Results not significant*

Both statements say the same thing, that is, if we use the accept H_0–reject H_0 format, then we must accept the supplier's claim ($\mu = .560$ mm), since we have no evidence to disprove the claim. My preference is to word the conclusion as follows:

Since the sample average of $\bar{x} = .553$ mm reasonably fits into the expected $\bar{x}$ distribution for $\mu = .560$ mm, we

Accept H_0: $\mu = .560$ mm

Now you might feel a little uncomfortable accepting H_0 since your sample average (.553 mm) did not fall precisely on the claimed population value of .560 mm. And at this point you might say, why don't we continue sampling to be more positive of our decision? Unfortunately, in most areas of research, further sampling is not practical. It is usually expensive, time-consuming, and in some cases physically impossible (when test circumstances cannot be duplicated). Certainly in this production control experiment, another random sample can be taken with relative ease, however in most studies in marketing, medicine, sociology, economics, and other fields, we often must rely on the results of one and only

*The words *not significant* have a very special meaning in statistical testing. They mean the results may reasonably be attributed to "chance fluctuation." In other words, $\bar{x}$'s may very well vary, fluctuate by chance, between .550 mm and .570 mm when $\mu = .560$ mm. Since we achieved an $\bar{x}$ (.553 mm) in this chance fluctuation range, we merely accept H_0. In broad terms, when sample results are,

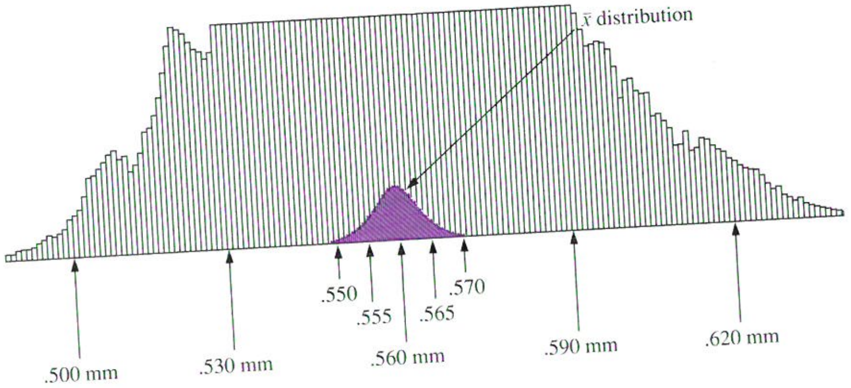
Not significant:	we accept H_0
Significant:	we reject H_0

one sample. Even in this production control experiment, no one wants to absorb the added time and expense of further sampling unless absolutely necessary. In other words, in statistical studies,

we normally base our decision on one and only one sample.

And we will conform to this practice in this text. So, to sum up our experiment, if our one sample average, $\bar{x}$, is reasonably close to the claimed μ, we accept H_0 as true and therefore accept the shipment of fiber-optic thread as meeting our specification of $\mu = .560$ mm.

However, this may cause some questions, such as: at what point do we grow suspicious that our sample $\bar{x}$ is *not reasonably close* to μ? For instance, what if our sample average turned out to be .550 mm or .540 mm or .577? Clearly, these values are on the very fringe of the ''expected'' sample averages. Observe:



In other words, at what value of $\bar{x}$ do we begin to grow suspicious that maybe the supplier's claim is false? Fortunately, there are certain industry standards that have proven reliable over decades of use. Although a number of industry standards exist, one of the most popular is the

Level of significance, $\alpha = 5\%$ (.05)

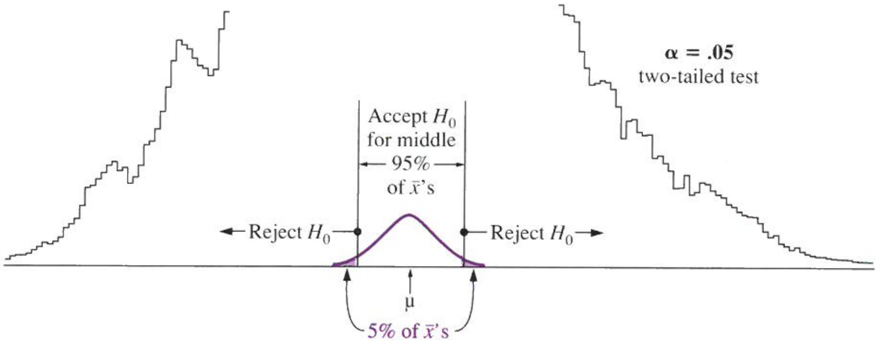
Although discussed in the last chapter, a brief review here might be helpful. Essentially, a level of significance sets up the cutoffs, or boundaries for accepting or rejecting H_0. For instance,

> For level of significance, $\alpha = 5\%$ (.05),* establish where the middle 95% of the $\bar{x}$'s are expected to fall if H_0 is true. Then, if the $\bar{x}$ you calculate

*Actually, many levels of significance are possible.

from your random sample falls inside (or exactly on the border) of this 95% range, accept H_0 as true. If the sample $\bar{x}$ falls outside, assume H_0 is false.

Visually we might present this $\alpha = .05$ hypothesis test as follows:



Two-Tailed Test

This is called a two-tailed hypothesis test since we have two tails of rejection (as shown shaded above). That is, we would reject the null hypothesis for any sample $\bar{x}$ falling in either of the *two* shaded tails.

To recap: if your sample $\bar{x}$ falls *inside* this 95% range (or on the border), accept H_0. If your sample $\bar{x}$ falls *outside* this range (that is, in the shaded tails), reject H_0. And this is precisely what is done in industry and research. Now let us repeat this problem as it would be worded and solved in practice.

Example

A supplier claims the average thickness (diameter) of its fiber-optic thread is .560 mm. You receive a shipment and decide to test their claim at a .05 level of significance by taking a sample of 36 randomly selected measurements, with the following results:

$$n = 36 \text{ measurements}$$
$$\bar{x} = .553 \text{ mm}$$
$$s = .030 \text{ mm}$$

What can we conclude?

Solution

A hypothesis test consists of three fundamental sequences as follows.

Sequence I. *Set up initial conditions:* H_0, H_1, and level of significance

		In Our Example, It Would Be
H_0:	State the null hypothesis, that is, the claim or assertion you wish to test.	H_0: $\mu = .560$ mm
H_1:	State the alternative hypothesis. In other words, if H_0 proves false, then what must we conclude?	H_1: $\mu \neq .560$ mm
α:	State the level of significance, α, that is, the risk of a Type I error (the risk of rejecting H_0 in error).	$\alpha = .05$ (5%)

Sequence II. *Assume H_0 true, use α to establish cutoffs* as follows:

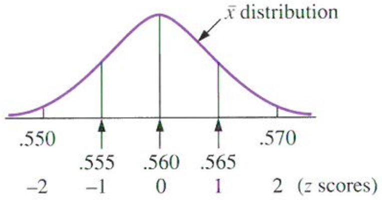
Calculate

$\sigma_{\bar{x}}$: We must remember we are dealing with $\bar{x}$'s and therefore must first calculate $\sigma_{\bar{x}}$, the standard deviation of the $\bar{x}$ distribution. Note in our formula for $\sigma_{\bar{x}}$, we used s (the standard deviation of the sample) as an estimator of σ (the population standard deviation).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} \approx \frac{.030}{\sqrt{36}} \approx \frac{.030}{6} \approx .005 \text{ mm}$$

Draw Curves

Using our above calculation, $\sigma_{\bar{x}} \approx .005$ mm, we estimate the spread of the $\bar{x}$ distribution.

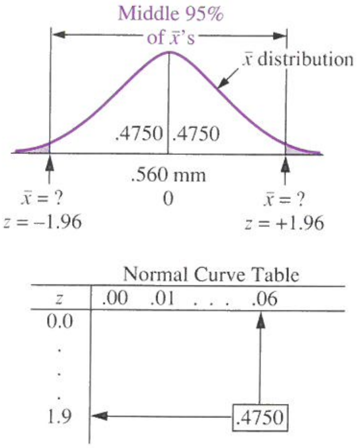


Establish Cutoffs (using α, the level of significance)

Our level of significance in this case is $\alpha = .05$ (5%), which in a two-tailed test implies we will accept the middle 95% of the

$\bar{x}$'s as our boundary for accepting H_0 as true. We now look up the z scores corresponding to the middle 95% of the $\bar{x}$'s, which turn out to be $z = -1.96$ and $z = +1.96$.

Remember: the normal curve table reads half the normal curve, starting from $z = 0$ out, so we look up $\frac{1}{2}$ of 95% or $47\frac{1}{2}$%, which in decimal form is .4750 (as shown at right).



Substituting the z scores of -1.96 and $+1.96$ into our formula, we solve for the $\bar{x}$ at the cutoffs.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$-1.96 = \frac{\bar{x} - .560}{.005}$$

Solving for $\bar{x}$:

$$\bar{x} = .550 \text{ mm}$$

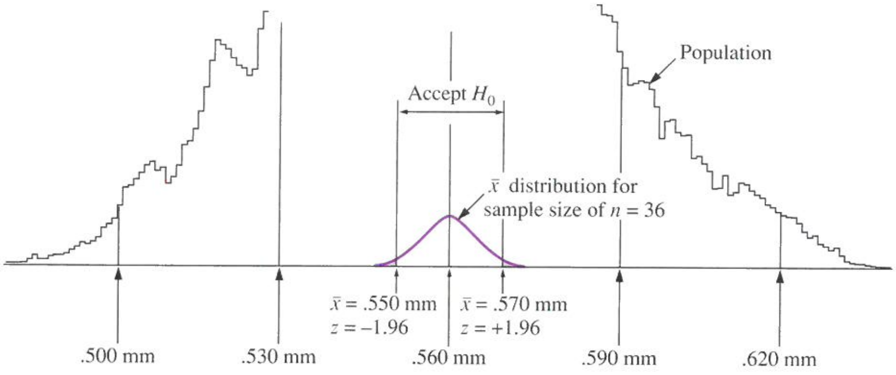
$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$+1.96 = \frac{\bar{x} - .560}{.005}$$

Solving for $\bar{x}$:

$$\bar{x} = .570 \text{ mm}$$

The completed solution would appear graphically as follows:



Note that the reject zones are shaded, that is, the zones where we would reject μ = .560 mm as being true. This is your risk of a Type I error (5%).

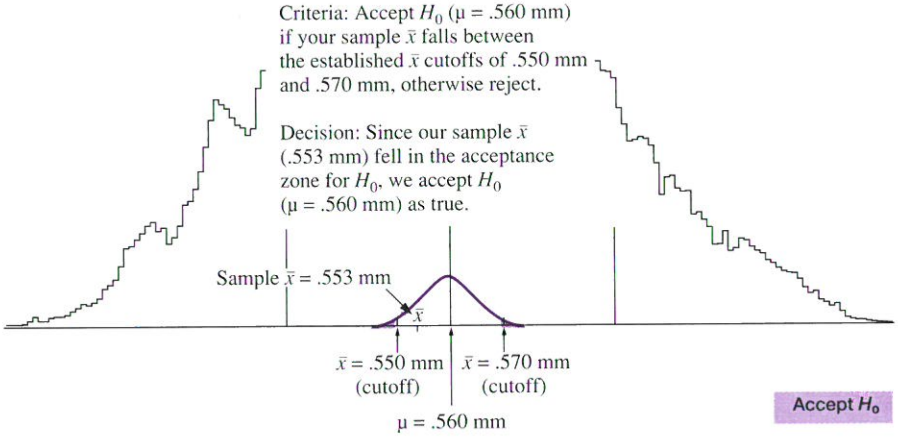
Sequence

III. *Accept or reject H_0 using your sample $\bar{x}$:* For this, two methods are available. Method One uses the actual value of the sample $\bar{x}$. Method Two uses the z score of the sample $\bar{x}$. Since each adds to understanding, we shall employ both.

METHOD ONE

This method uses the actual value of the sample $\bar{x}$ (.553) in the decision-making process.

Recall: Our sample results were as follows:
n = 36 measurements
$\bar{x}$ = .553 mm



Criteria: Accept H_0 (μ = .560 mm) if your sample $\bar{x}$ falls between the established $\bar{x}$ cutoffs of .550 mm and .570 mm, otherwise reject.

Decision: Since our sample $\bar{x}$ (.553 mm) fell in the acceptance zone for H_0, we accept H_0 (μ = .560 mm) as true.

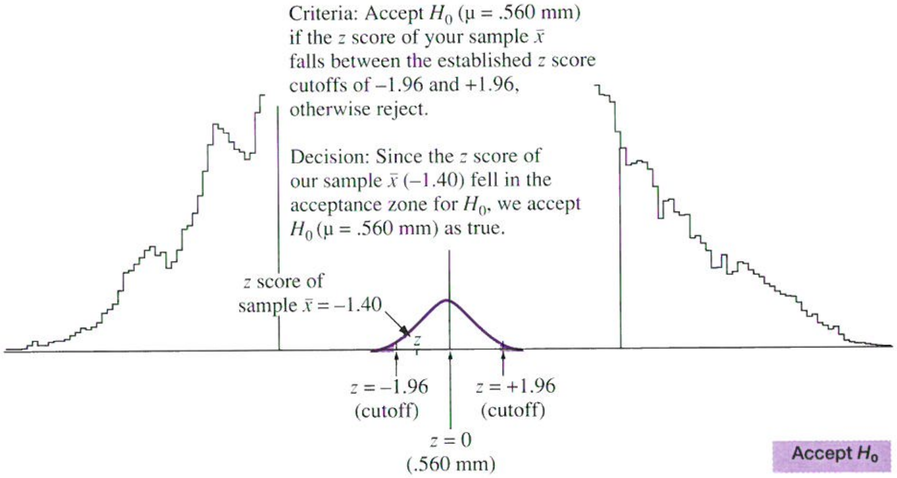
Sample $\bar{x}$ = .553 mm
$\bar{x}$
$\bar{x}$ = .550 mm (cutoff)
$\bar{x}$ = .570 mm (cutoff)
μ = .560 mm
Accept H_0

METHOD TWO

This method uses the z score of the sample $\bar{x}$ in the decision-making process. To use this method, however, we must first calculate the z score of our sample $\bar{x}$ (.553 mm), as follows.

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{.553 - .560}{.005}$$

$$z = -1.40$$



Criteria: Accept H_0 (μ = .560 mm) if the z score of your sample $\bar{x}$ falls between the established z score cutoffs of −1.96 and +1.96, otherwise reject.

Decision: Since the z score of our sample $\bar{x}$ (−1.40) fell in the acceptance zone for H_0, we accept H_0 (μ = .560 mm) as true.

Whether we use the actual value of the sample $\bar{x}$ or the z score of the sample $\bar{x}$, we will always make the same decision. In this case, we accept H_0. Generally, the z score is preferred by those most familiar with statistical technique since the z score is a more informative measure. Note that we better understand the position of the sample $\bar{x}$ if we say it is -1.40 standard deviations from the claimed μ than if we merely presented its actual value of .553 mm.*

Answer

The final answer may be presented in a number of ways, depending on the technical expertise of those reading the report.

a. If the report is to be presented to individuals unfamiliar with statistical technique, perhaps the following offers a clear approach:

Since the sample average we obtained from the shipment (.553 mm) falls inside the range (.550 to .570) where we would most likely expect sample averages to fall if H_0 were true, we accept H_0: $\mu = .560$ mm, and therefore accept the shipment.

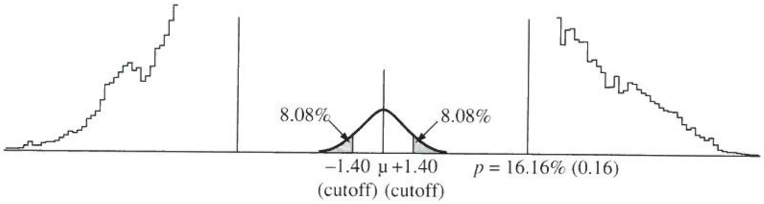
Accept H_0

b. However, this same answer may very well appear in a technical report worded in terms of z scores, as follows:

Since our sample z of -1.40 is not less than -1.96, the null hypothesis cannot be rejected. The difference between .553 mm and .560 mm is not large enough to provide evidence at the .05 level of significance that the shipment does not meet supplier's specification.

Null hypothesis cannot be rejected

**P*-value approach: Actually a third method is also used. This method calculates the probability of achieving a result *at least* as many standard deviations from the expected value as your sample result.



Let's reconsider the above example. Since we achieved a sample result 1.40 standard deviations from the expected value, μ (calculated in Method Two), we shade all the area that is *at least* 1.40 standard deviations from μ. Note in a two-tailed test, we shade *both* tails. Next we look up the probability of achieving a sample result in this shaded area, which is 16.16% (8.08% in each tail). This is our p-value. This is usually expressed in technical reports and computer software printouts as either $p = .16$ or $p > .05$ (meaning the probability of achieving this sample $\bar{x}$ is greater than the α level of the test).

For $p \geq \alpha$, Accept H_0, otherwise reject

Since in our case, $.16 \geq .05$, we Accept H_0.

c. Then again, many reports simply present the results as

$z = -1.40$ (not significant).

Results not significant*

Believe it or not, all three answers say the same thing. Try to understand the technical explanations using z scores, since this is typical of how research reports are presented. ■

Control Charts

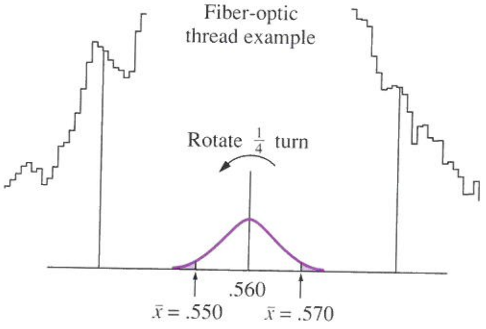
In production studies and occasionally in marketing, medical, and other studies, the same hypothesis test may be repeated a number of times. For instance, what if this telephone cable manufacturer in the prior problem were to accept this shipment of fiber-optic thread and then ordered additional fiber-optic thread under the same specifications, to be delivered once a month for several months? Each monthly shipment may very well be tested in an identical manner. When essentially the same test must be repeated on a periodic basis, a **control chart** can be set up as follows:

Construction of Control Chart

1. On a graph, establish cutoffs for a given hypothesis test. In industrial production, cutoffs are usually referred to as *control limits*.
2. Rotate graph $\frac{1}{4}$ turn counterclockwise, extending the cutoff lines to the right. Shade rejection zone.
3. Plot each sample $\bar{x}$ sequentially to the right. Connect each $\bar{x}$ to prior result with a line segment.

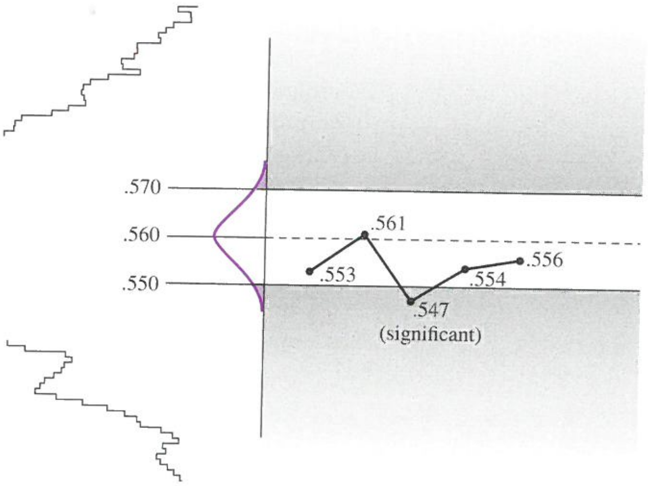
In a control chart, you may choose to use either actual values or z scores to represent the readings. For instance, say we use actual values, we would proceed (using our fiber-optic thread example) as follows.

*Again, the words *not significant* have a very special meaning in statistical testing. Essentially, *not significant* means: the sample results (in this case, $\bar{x} = .553$ or $z = -1.40$) are considered ''chance fluctuation.'' In other words, we would expect to find $\bar{x}$'s between ±1.96 standard deviations of the mean if H_0 were true. Since the z score (-1.40) of our sample $\bar{x}$ was in this chance fluctuation range between ±1.96 standard deviations, it is deemed not significant and we accept H_0.



Cutoffs established, taken from prior example.

Rotate $\frac{1}{4}$ turn counterclockwise, extending cutoff lines to the right and shading rejection zone (as shown in next diagram).



Now let's say we receive 5 shipments over several months and calculate the sample $\bar{x}$ for each as follows.

$\bar{x} = .553$ mm
$\bar{x} = .561$ mm
$\bar{x} = .547$ mm (significant)
$\bar{x} = .554$ mm
$\bar{x} = .556$ mm

Each sample $\bar{x}$ is plotted sequentially as the shipment comes in and connected with a line segment to prior result (as shown above).

Note that one sample $\bar{x}$ (.547 mm) was marked ''significant.'' This means, based on this one sample average, we would reject this particular shipment as not meeting specifications. At this point, the production supervisor would likely be called in. After verifying results, the supervisor may very well call the manufacturer of the fiber-optic thread to inform them that their process was not meeting specification, and most likely ''out of control.'' A process is deemed out of control when sample $\bar{x}$'s fall outside the control limits for acceptance of H_0 and we suspect a possible deterioration of the process.

Note that a control chart provides a clear visual history of this hypothesis test. Often we learn more about a process by keeping this kind of record. Sometimes we can spot a trend, a process going out of control *before* a significant sample $\bar{x}$ is achieved. Or we may be able to pick up slight shifts in the value of μ, even though sample $\bar{x}$'s are in control. For a process in control, the sample $\bar{x}$'s should fluctuate (usually in a ragged pattern) around the value of μ. Notice that the $\bar{x}$'s we calculated, .553, .561, .547, .554, and .556, seem to fluctuate more around the value of .555 (than the value .560). If this trend continues for future shipments, we may very well suspect the thickness of the fiber-optic thread shipped may be on average, $\mu = .555$ mm. Of course, whether or not this slight shift makes a difference in our production would have to be assessed.

A *control chart*[*] provides a clear visual history of a repetitive test.

7.2 One-Tailed Hypothesis Tests (Large Sample, $n \geq 30$)

A one-tailed hypothesis test is quite similar in method to a two-tailed hypothesis test, except in a one-tailed test, the Type I error risk (α) is assigned to only *one* tail of the $\bar{x}$ distribution.

One-Tailed Hypothesis Test[†]

All the Type I error risk, α, is assigned to *one* tail of the $\bar{x}$ distribution, and we reject H_0 for any sample $\bar{x}$ falling in this *one* tail only.

The α risk may be assigned to either the right or left tail, depending on the hypothesis you wish to test. The following two examples demonstrate this.

[*] Historical note: Walter Shewhart first developed control charts in 1924, which were tested and developed within the Bell Telephone System, 1926–1931. For further historical reading on this topic, refer to, W. Peters, *Counting for Something* (New York: Springer-Verlag, 1987), Chapter 16, ''Quality Control,'' pp. 151–162.

[†] Actually, some controversy surrounds the use of one-tailed hypothesis testing. Refer to D. Howell, *Statistical Methods for Psychology* (Boston: PWS Publishers, 1982, pp. 64–66) for a discussion of one- and two-tailed tests. Essentially, Howell argues that an investigator may start with a one-tailed test, yet reject in two tails, thus inadvertently increasing the α level of the experiment. Howell also states, ''A number of empirical studies have shown that the common statistical tests . . . are remarkably robust when they are run as two-tailed tests, but are not always so robust when run as one-tailed tests.'' **Robustness** is the degree to which you can violate the assumptions of a test and yet leave the validity more or less unaffected.