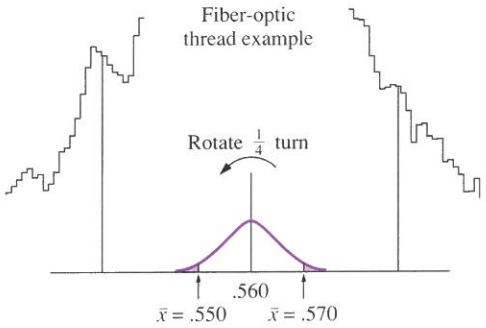
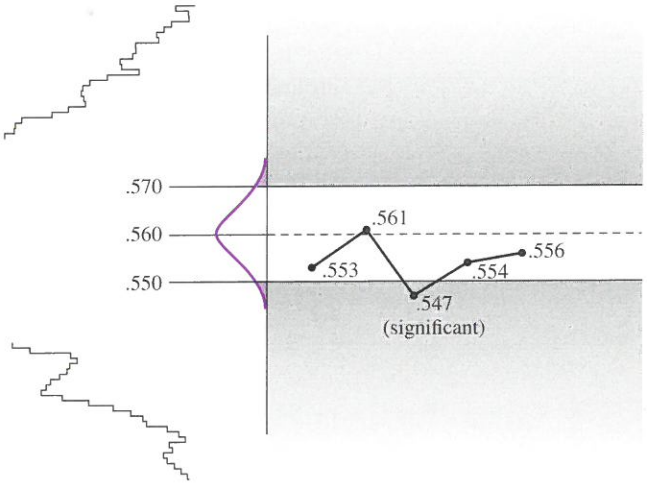


Cutoffs established, taken from prior example.

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

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†

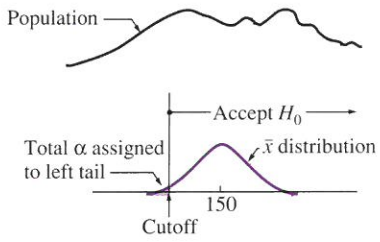
All the Type I error risk, α, is assigned to *one* tail of the $\bar{x}$ distribution, and we reject H_0 for any sample $\bar{x}$ falling in this *one* tail only.

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.

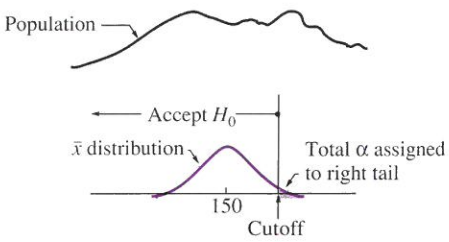
*Historical note: Walter Shewhart first developed control charts in 1924, which were tested and developed within the Bell Telephone System, 1926–1931. For further historical reading on this topic, refer to, W. Peters, *Counting for Something* (New York: Springer-Verlag, 1987), Chapter 16, "Quality Control,"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

To Test the Hypothesis
H_0: $\mu = 150$ *or more,* we would assign the total α risk to the left tail in the $\bar{x}$ distribution and reject H_0 for any sample $\bar{x}$ in this left tail, as shown shaded below.



To Test the Hypothesis
H_0: $\mu = 150$ *or less,* we would assign the total α risk to the right tail in the $\bar{x}$ distribution and reject H_0 for any sample $\bar{x}$ in this right tail, as shown shaded below.



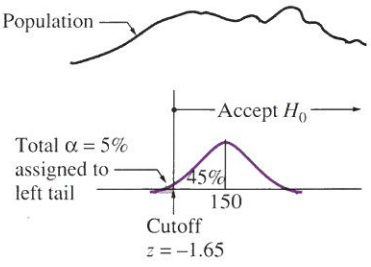
Notice in each example above, the total Type I error risk, α, was assigned to only *one* tail in the $\bar{x}$ distribution. This will affect the determination of the z score at the cutoff. Other than this, a one-tailed test is conducted in almost an identical manner as a two-tailed test. For instance, suppose we wish to test the following null hypothesis:

$$H_0: \mu = 150 \text{ or more}$$

at an $\alpha = .05$ level of significance; we would proceed as follows:

First, assign the total α risk (.05 or 5%) to the left tail of the $\bar{x}$ distribution. Why the left tail? Well, because we reject H_0 only if our sample average falls *significantly below* 150. (Note: you would not reject the hypothesis, $\mu = 150$ or more, for a sample $\bar{x}$ greater than 150.)

Second, to find the z score at the cutoff, we look up in the normal curve table, 45% (50% − 5% = 45%). Remember, the table reads from the center of the normal curve out. (Note: .4500 falls midway between .4495 and .4505 in the table, thus, round to the higher number, .4505, which is $z = 1.65$.) Since the cutoff is *below* μ, we apply a negative sign to the z score; thus $z_{\text{cutoff}} = -1.65$.

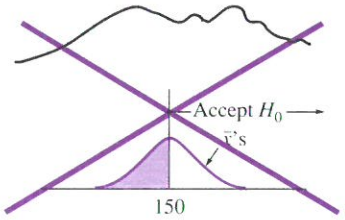


Normal Curve Table

z	.00	.01	.02	. . .	.05
0.0					
.					
.					
.					
1.6					.4505

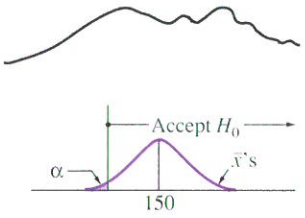
The decision-making process, in this case, is: Accept H_0 for any sample $\bar{x}$ to the right of the cutoff, otherwise reject. That is, we reject H_0 for any sample $\bar{x}$ in the shaded tail.

At this point you might ask, why don't we make sure μ is 150 or more by shading all of the values below 150 as follows?



This is wrong

Remember, we are talking about sample averages, $\bar{x}$'s, and $\bar{x}$'s tend to fluctuate around the population average, μ. That is, μ may very well be exactly 150 and still you could get sample $\bar{x}$'s *below* this value. So, we must leave some margin below μ for the $\bar{x}$'s to fluctuate, as follows:



Correct for
$\mu = 150$ or more

To recap: in a one-tailed hypothesis test, you assign the total α risk to *one* tail of the $\bar{x}$ distribution (which we shade), and you reject H_0 for any sample $\bar{x}$ in this shaded tail. Other than this, a one-tailed hypothesis test is conducted in almost an identical manner as a two-tailed hypothesis test.

At this point, I have found the following two reminders helpful.

Keep in mind for all hypothesis tests

1. Use α, the level of significance, to establish the cutoff(s), and
2. Shade where you would reject H_0.

Applications

Now let's see how this works in a study in psychiatric medicine.

Example

Elavil,* a powerful sedating drug prescribed by psychiatrists, has been proven effective over decades of use in the treatment of depression, however Elavil can have side effects (causing high blood pressure, dry mouth and impotence problems, blurred vision, etc.), so daily dosages must be minimized. However, one patient may need 75 mg per day to relieve depression while another patient may need 250 mg per day to have the same effect, depending on the individual.

Suppose a leading trade journal makes the claim that the average effective dosage nationwide for patients is *at least* 150 mg per day. Concerned such an article may influence psychiatrists to unnecessarily increase dosages, suppose the National Institute of Mental Health in Bethesda, Md., conducts a test by randomly sampling 400 patients nationwide, with the following result:

n = 400 patients
$\bar{x}$ = 141.6 mg/day minimum effective dosage
s = 48.2 mg/day

Use this sample result to test at a level of significance of $\alpha = .03$, the trade journal's claim

μ is *at least* 150 mg/day.

Solution

This is a one-tailed hypothesis test since in effect the trade journal's claim is $\mu = 150$ mg/day *or more*. In other words, you would reject the claim only if your sample $\bar{x}$ was unreasonably *below* 150 mg/day. That is, we reject only in one direction.

A hypothesis test consists of three fundamental sequences as follows:

Sequence

I. Set up initial conditions

State null hypothesis, the initial claim you wish to test:	H_0: μ = 150 mg/day or more ($\mu \geq 150$)
State alternative hypothesis. If H_0 proves false, what must we conclude?	H_1: μ is less than 150 mg/day ($\mu < 150$)
State the risk of rejecting H_0 in error, the level of significance, α.	α = .03 (3%)

*Elavil is part of the tricyclic family of antidepressant drugs, along with Sinequan, Tofranil, and Norpramin. Each relieves depression with varying degrees of sedating effect. Elavil is one of the more powerful sedating drugs, often used when agitation or sleeplessness accompany depression.

Sequence II. Assume H_0 true, use α to establish cutoff(s)

Calculate

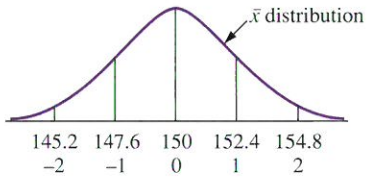
$\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} \approx \frac{48.2}{\sqrt{400}} \approx \frac{48.2}{20} \approx 2.41$$

(Note s was used to estimate σ.)

Draw curves

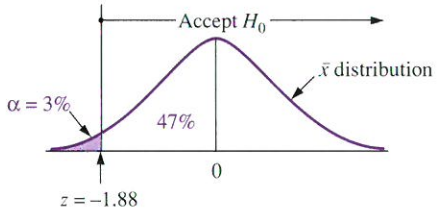
Using our above calculation, $\sigma_{\bar{x}} = 2.4$, we estimate the spread of the $\bar{x}$ distribution.



Establish cutoffs using α

Assign the total α risk (.03 or 3%) to the *left* tail of the $\bar{x}$ distribution because we would reject H_0 only if our sample $\bar{x}$ falls significantly below 150 mg/day.

Next, to find the z score at the cutoff, we look up 47% in the normal curve table (50% − 3% = 47%). Remember, the table reads from the center of the normal curve out. (47% in decimal form is .4700; the closest value is .4699.)



Normal Curve Table

z	.00	.01	.02	. . .	.08
0.0					
.					
.					
.					
1.8					.4699

Since the cutoff is *below* $\mu = 150$, the z score will be negative. Thus, $z = -1.88$. Substituting the z score of -1.88 into our formula, we solve for the $\bar{x}$ value at the cutoff.

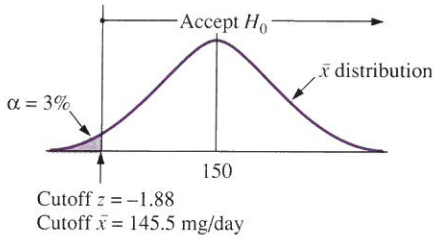
$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$-1.88 = \frac{\bar{x} - 150}{2.4}$$

Solving for $\bar{x}$:

$$\bar{x} = 145.5 \text{ mg/day at the cutoff}$$

Thus, the values at the cutoff might be represented as follows:



Note that in the diagram above there is no indication of the population, such as population standard deviation lines. This is because the sample size is so large ($n = 400$), causing the $\bar{x}$'s to cluster so tightly around μ that the population standard deviation lines are far out of view.

Sequence

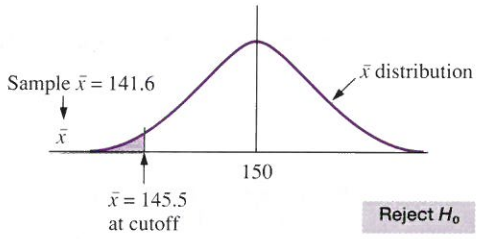
III. Accept or reject H_0 using your sample $\bar{x}$

METHOD ONE
This method uses the actual value of the sample $\bar{x}$ (141.6 mg/day) in the decision-making process.

Criteria: Accept H_0 ($\mu = 150$ mg/day or more) if the sample $\bar{x}$ falls *above* (or on border of) the $\bar{x}$ cutoff of 145.5 mg/day, otherwise reject.

RECALL: our sample results were as follows:
$n = 400$ patients
$\bar{x} = 141.6$ mg/day.

Decision: Since our sample $\bar{x}$ (141.6 mg/day) fell in the rejection zone, we reject H_0 and accept H_1, the **alternative hypothesis** (μ is *less than* 150 mg/day).



METHOD TWO

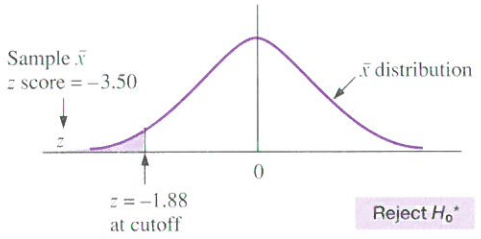
This method uses the z score of the sample $\bar{x}$ in the decision-making process. To use this method, however, we must first calculate the z score of our sample $\bar{x}$ (141.6 mg/day), as follows:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{141.6 - 150}{2.4}$$

$$z = -3.50$$

Criteria: Accept H_0 (μ = 150 mg/day or more) if the z score of the sample $\bar{x}$ falls *above* (or on border of) the z score cutoff of -1.88, otherwise reject.

Decision: Since the z score of our sample $\bar{x}$ (-3.50) fell in the rejection zone, we reject H_0 and accept H_1, the alternative hypothesis (μ is *less than* 150 mg/day).



Whether we use the actual value or z score of the sample $\bar{x}$, we will always make the same decision. In this case, we reject H_0. This implies we accept H_1 (μ is less than 150 mg/day).

Answer

The final answer may be presented using actual values or z scores. We will use both.

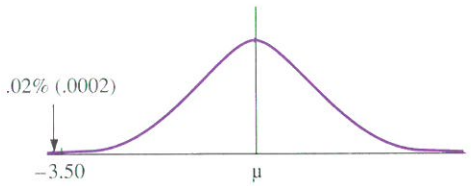
Actual values — Since the sample average (141.6 mg/day) was below the cutoff of 145.5 mg/day, we reject H_0. Therefore we

Accept H_1: μ is *less than* 150 mg/day

z scores — Since the sample average z score (-3.50) was below the cutoff of -1.88, we reject H_0. Therefore we

Accept H_1: μ is *less than* 150 mg/day

*P-value approach: Actually, as mentioned in the prior section, 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 consider this using the above example. Since we achieved a sample result of -3.50 standard deviations from the expected value, μ, we shade all the area that is *at least* -3.50 standard deviations from μ. Note in a one-tailed test, we shade in *one* tail only. Next we look up the probability of achieving a sample result in this shaded area, which is .02% (.0002), a negligible amount. This is our p-value. This is usually expressed in research reports and computer software printouts as either p = .0002 or $p < .03$ (meaning the probability of achieving this sample $\bar{x}$ is less than the α level of the test).



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

Since in our case, .0002 < .03, we reject H_0.

Other ways the answer may be expressed:

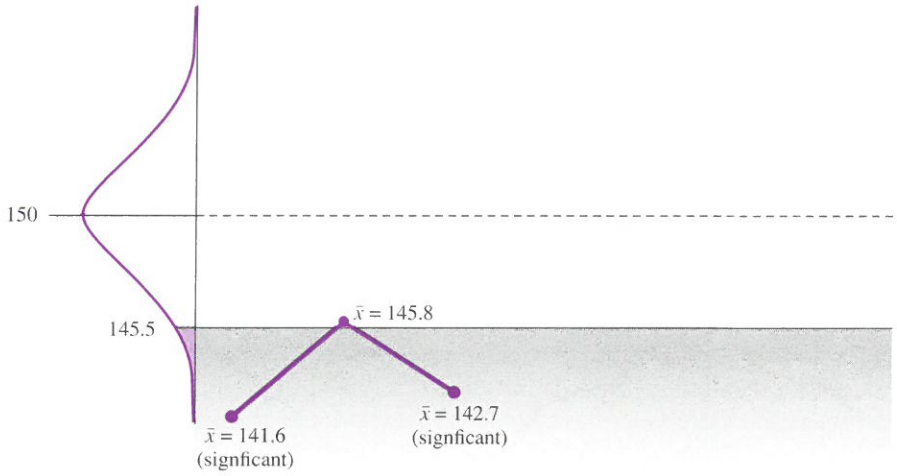
The null hypothesis is rejected,

or

$z = -3.50$ (significant) ■

Control Charts

Besides this sample result, $\bar{x} = 141.6$ mg, as presented in the previous example, suppose two identical studies (same hypothesis, same sample size and level of significance) yielded $\bar{x} = 145.8$ mg and $\bar{x} = 142.7$ mg. Plot these three results into a *control chart* and indicate significant findings.

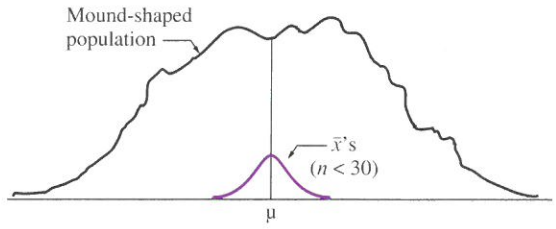


7.3 Small-Sample Hypothesis Tests ($n < 30$)

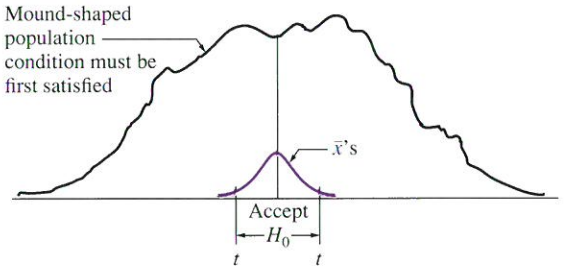
In statistical testing, small-sample sizes ($n < 30$) may also be used effectively, but only when the following conditions are satisfied.

When using small samples ($n < 30$)

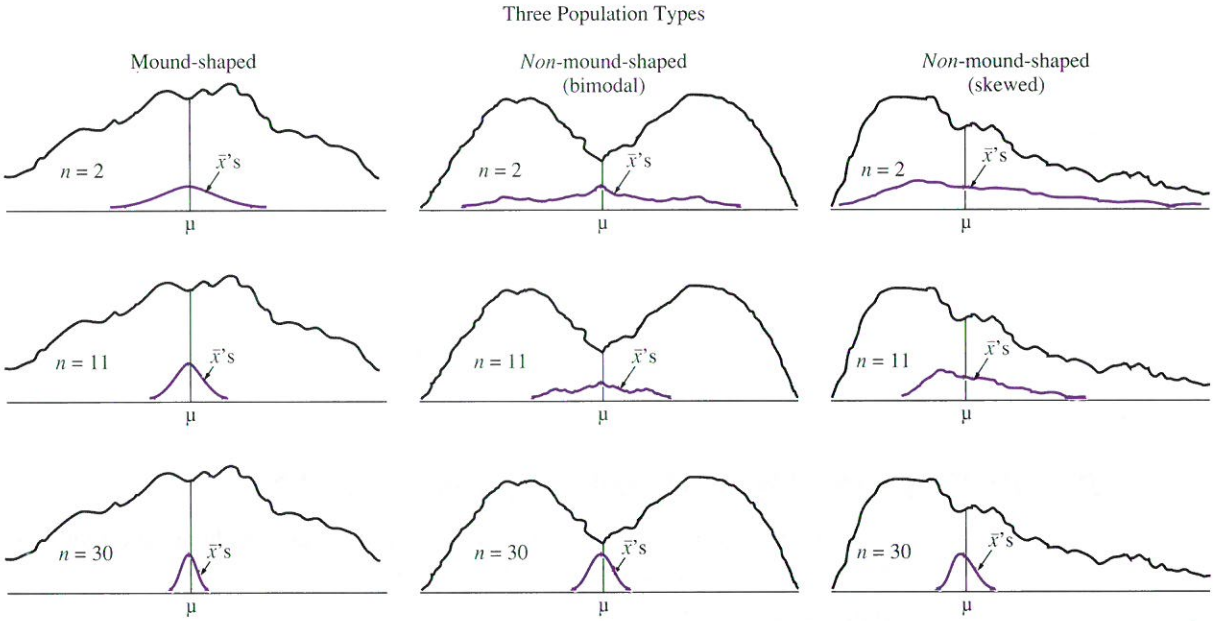
1. Your population should be normally distributed or at least somewhat mound shaped, *and*



2. If we use s to estimate σ in the calculation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, we must use a t score, not a z score, to define the number of standard deviations the $\bar{x}$'s would be expected to fall from μ.



Essentially, the first condition (population normal or at least mound shaped) must be satisfied to ensure that the $\bar{x}$'s cluster close enough to μ in a reasonably normal shape such that accurate predictions can be made. When your population is *not* mound shaped, the $\bar{x}$'s spread out in a variety of patterns, often quite far from μ, as illustrated below:

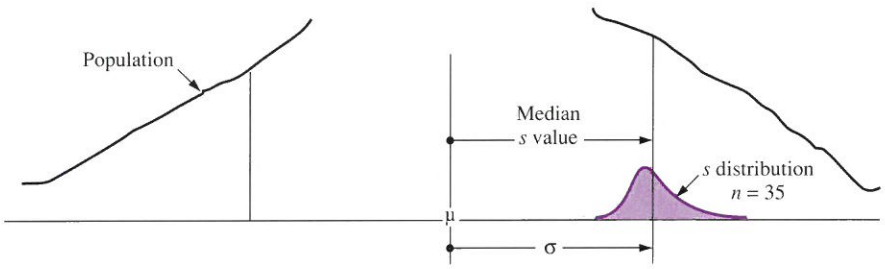


Notice it is only in the mound-shaped population that the $\bar{x}$'s cluster close to μ in a normal distribution for *all* sample sizes. For this reason, small-sample sizes can be used when your population is mound shaped. Notice in *non*-mound-

shaped populations, a small-sample size ($n < 30$) will often produce $\bar{x}$'s that are quite far from μ, forming a variety of *non*normal patterns, making predictions about where the $\bar{x}$'s will fall quite perilous. Thus, when sampling from *non*-mound-shaped populations, small-sample sizes should be avoided. Of course, when your sample size grows sufficiently large (usually $n \geq 30$ is considered sufficiently large), the $\bar{x}$'s draw in quite close to μ for almost any shaped population, even for *non*-mound-shaped populations. Thus, for $n \geq 30$, the population shape has little effect on the $\bar{x}$ distribution shape; the $\bar{x}$'s will distribute normally about μ for nearly any shaped population, thus assuring reliable predictions.*

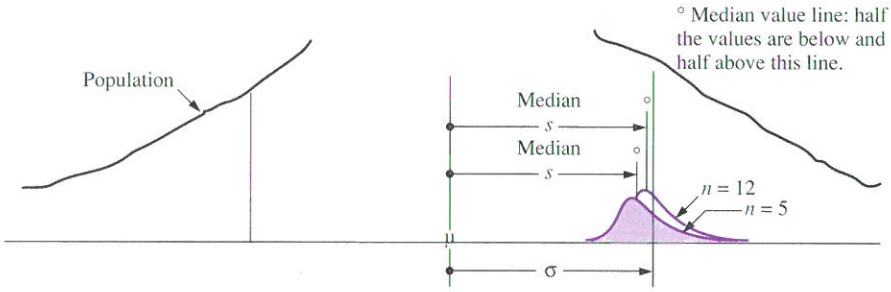
Assuming we have a mound-shaped population, a second condition must also be satisfied. If we chose to use s to estimate σ (which we most often do) in the formula, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, this necessitates a correction factor in our calculations—which we shall call the t score adjustment. Although this is more fully discussed in chapter 8, section 8.4, an overview here might be helpful.

Whereas, a large sample s is a reasonably good estimator of σ, meaning that s-values cluster quite close to σ. More specifically, if we were to take thousands of samples of size, say, $n = 35$ and calculated the standard deviation, s, for each sample and plotted these thousands of s's, the resulting distribution would be clustered relatively close around the true value of σ, as follows:



*Technical note: For *highly* unusual population shapes (such as, a population with an extraordinary skew), n may have to be larger than 30 to be assured of a normally distributed $\bar{x}$ distribution. An example of this might be the distribution of annual salaries of workers in lower Manhattan, which includes the highly skewed million-dollar-plus salaries of many Wall Street executives.

This is not the case for small-sample s's. Small-sample s's tend to *under*estimate σ; in fact, small-sample s's tend to underestimate σ more and more as n (the sample size) decreases,* illustrated as follows:

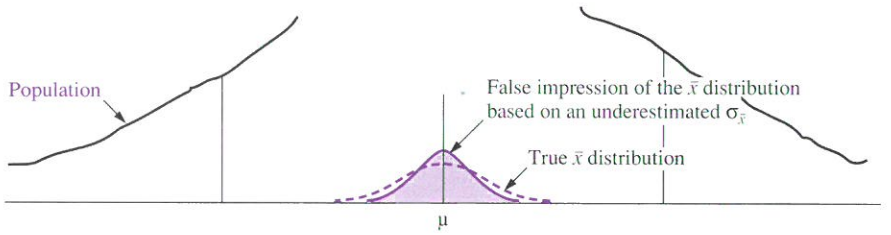


Notice that as your sample size decreases ($n = 35$, $n = 12$, $n = 5$), the s's tend to shift to lesser values, such that the median s value falls farther and farther below the true value of σ. This means, more often than not, when you calculate the s of your sample, it will be *less* in value than σ.

Now, since we use s in place of σ in the formula,

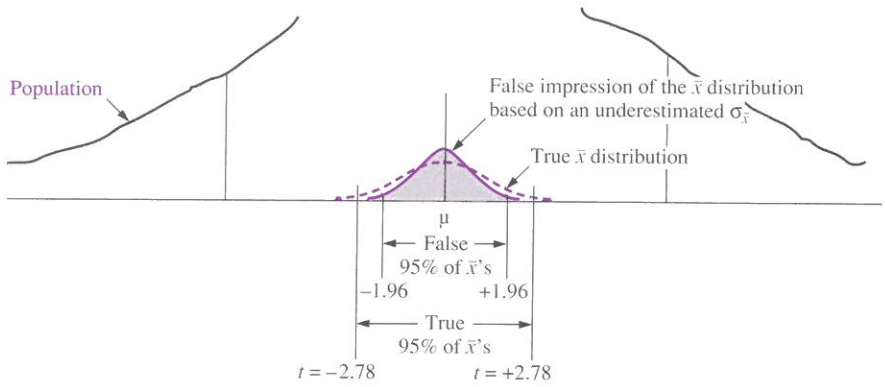
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

If s underestimates σ, then $\sigma_{\bar{x}}$ will also be *under*estimated. In other words, more often than not, we will calculate a $\sigma_{\bar{x}}$ that is less in value than it actually is—because of the *under*estimated s. Visually, this might be represented as follows:



*Technical note: s^2 distributes around σ^2 in a chi-square-shaped distribution and on average $s^2 = \sigma^2$ for all sample sizes, however the median s^2 value drops below σ^2 as n decreases. The distribution of s is similar and can be found by compressing the base line suitably, to paraphrase W. Gossett (the statistician who originally published these findings in *Biometrika*, VI. pp. 1–25, 1908, under the pen name, Student), the distribution of s^2 has a direct linear relationship to the distribution of χ^2, chi-square, specifically, $s^2 = (\frac{\sigma^2}{n-1})\chi^2$.

Fortunately, we can compensate for this. Say, for instance, we conduct a two-tailed hypothesis test at $\alpha = .05$, using a sample size of $n = 5$. We know, for a large sample, $\alpha = .05$ implies an interval bounded by ± 1.96 standard deviations. In other words, 95% of the $\bar{x}$'s are expected to fall within ± 1.96 standard deviations of μ. However, in the case of $n = 5$, we must open up the interval to ± 2.78 standard deviations, use the letter t and say, 95% of the $\bar{x}$'s will fall in the interval between ± 2.78 standard deviations, illustrated as follows:



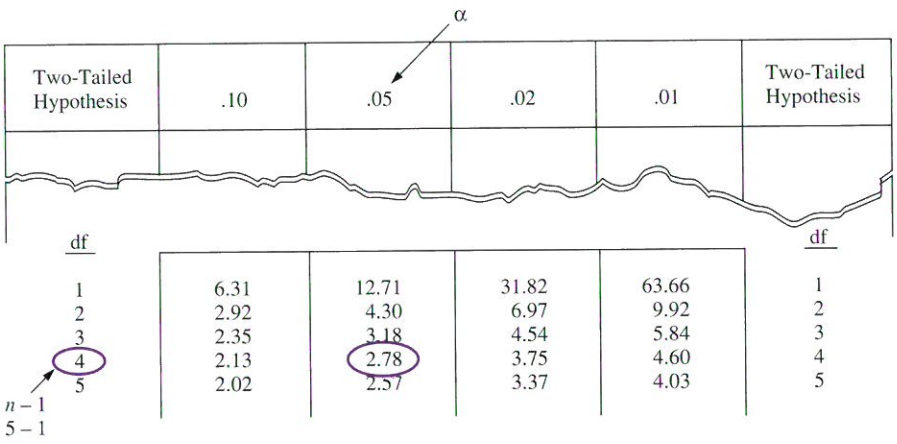
In other words, to ensure we have enclosed the *true* 95% of the $\bar{x}$'s when using small sample size $n = 5$, we must go out farther, to ± 2.78 standard deviations (*not* ± 1.96). And instead of using the letter z to represent the number of standard deviations, we use the letter,* t.

Of course, at this point, you might ask, where do we locate this t score adjustment of ± 2.78? The answer is simple; we look it up in the t tables in the

*Technical note: Actually, a number of liberties were taken with this explanation. In reality, t-values were derived empirically by W. Gossett. He started with a known mound-shaped population of values, then literally took thousands of random samples of size $n = 2$. For each, he calculated the number of standard deviations (z score) the sample $\bar{x}$ *appeared* to fall from μ (using the s of the sample for each calculation). These z scores (many of which were distorted because of the *under*estimations of $\sigma_{\bar{x}}$) were plotted into a histogram that he called the t distribution for $n = 2$. This distribution resembles a normal curve, but is more flat on top and spread out in the tails. He repeated this process for $n = 3$, $n = 4$, etc. In the case above, for instance, where $n = 5$, he noted 95% of the $\bar{x}$'s appear to fall within ± 2.78 standard deviations of μ, based on a number of distorted estimates of $\sigma_{\bar{x}}$. He had also derived equations for the distributions from fundamental theory and used these empirical results to validate these equations. (For further details, refer to endnote 1.)

Further technical note: W. Gossett in his original 1908 *Biometrika* article claimed the population shape can deviate quite far from normal before this would influence the predictive value of these t distributions.

back of the text. Since we are conducting a two-tailed hypothesis test at $\alpha = .05$, we look under *that* particular column, then down to degrees of freedom (df) of 4 (our sample size minus one, $5 - 1 = 4$), as follows:



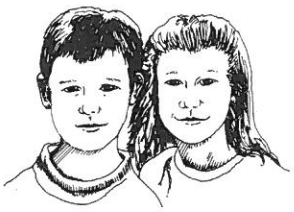
Degrees of freedom (or df): Although difficult to define completely,* let us say for our purposes, df defines a value that allows us to identify the correct sampling distribution and thus the proper cutoff value(s) for a given experiment. Essentially each df value represents a different sampling distribution.

Applications

Now let's see how small-sample testing works in two experiments. The first is a classroom experiment concerning social intelligence in children.

*Technical note: df is more formally defined as: "For any problem the number of degrees of freedom is the number of variables reduced by the number of independent restrictions on those variables," H. Walker and J. Lev, *Elementary Statistical Methods*. (New York: Holt, Rinehart, and Winston, 1969), p. 276. The concept was known to Carl Gauss (1826) but it wasn't until Sir Ronald Fisher's 1915 paper, "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, Vol. x, pp. 507–521, where Fisher applied *n*-dimensional geometry to its application that it gained widespread attention. The concept is quite advanced, but for those who wish a deeper understanding, refer to H. Walker, "Degrees of Freedom," *Journal of Educational Psychology* 31 (1940): pp. 253–269.

Example



Social IQ in young children, that is, a child's ability to properly assess nonverbal messages of teachers and peers (such as, tone of voice, body gesture, and facial expression) and properly assess social boundaries (such as, sensing how close to stand, responding at proper intervals, and smoothness in entering groups and in communicating emotions like anger and happiness) may be a more accurate measure than mental IQ in predicting later academic achievement and overall success throughout life, according to researchers at Harvard, University of Illinois, University of North Carolina, and other institutions.

Suppose Ms. Peach has her Lake County, Illinois, elementary school class of 9 students (underachiever class) tested on a social IQ scale, with the following results.

$n = 9$ underachiever students
$\bar{x} = 85.3$ social IQ
$s = 11.7$ (assume a normal population)

a. Test the hypothesis at $\alpha = .02$ that these students came from a population with social IQ, $\mu = 100$ (which is average for children of this age). Are the results significant?

b. If the data constitutes a valid random sample of underachiever students, what conclusions can be drawn? Briefly discuss validity.

Solution

Since a small-sample size, $n = 9$, was used, we must be careful that two conditions are first satisfied: (1) the population is at least somewhat mound shaped (it was stated, assume a normal population, so this condition is satisfied) and (2) if s is used to estimate σ, t scores must be used, not z scores (notice: we were *not* given σ, so s must be used to estimate σ, thus we must use t scores).

A hypothesis test consists of three fundamental sequences as follows.

Sequence I. Set up initial conditions

State null hypothesis, the initial claim you wish to test.	H_0: $\mu = 100$ social IQ
State the alternative hypothesis. If H_0 proves false, what must we conclude?	H_1: $\mu \neq 100$ social IQ
State the risk of rejecting H_0 in error, the level of significance, σ.	$\alpha = .02$ (2%)

Sequence II. Assume H_0 true, use α to establish cutoff(s)

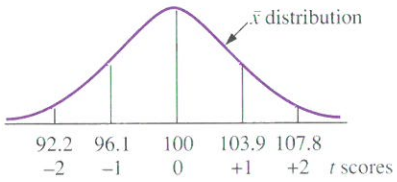
Calculate

$\sigma_{\bar{x}}$:
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} \approx \frac{11.7}{\sqrt{9}} \approx \frac{11.7}{3} \approx 3.9$$

(Note s was used to estimate σ above; a small-sample s will tend to underestimate σ, causing $\sigma_{\bar{x}}$ to be underestimated, too.)

Draw Curves

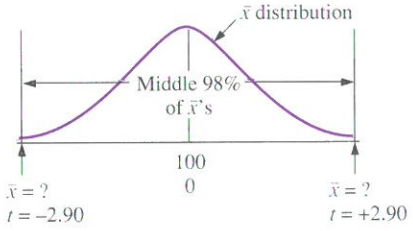
Using our above calculation, $\sigma_{\bar{x}} \approx 3.9$, we estimate the spread of the $\bar{x}$ distribution. (Keep in mind, $\sigma_{\bar{x}}$ is probably *under*estimated, however we will compensate for this by using t scores.)

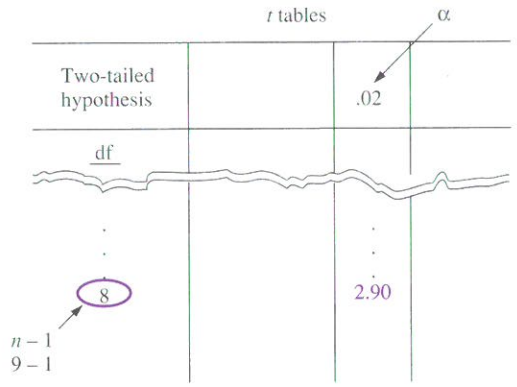


Establish Cutoffs using α

Our level of significance in this case is $\alpha = .02$ (2%) which in a two-tailed test implies we will accept the middle 98% of the $\bar{x}$'s as our boundary for accepting H_0 as true. In our t tables, we look up: two-tailed hypothesis, $\alpha = .02$, to obtain $t = \pm 2.90$ standard deviations. Remember: Look down the df (degrees of freedom) column to 8 (df $= n - 1 = 9 - 1 = 8$).

Note: Had this been a *large* sample, the boundaries would have been ±2.33 standard deviations instead of ±2.90.





We solve using the z formula, only now we use the letter t in place of z. We can view the t score* as an adjustment to the z score (to compensate for the underestimations of $\sigma_{\bar{x}}$).

*Technical note: Actually, we are sampling from the t distribution (for $n = 9$), which is, essentially, a distribution of distorted z values. More specifically, it is a distribution of thousands and thousands of z scores describing how far in standard deviations the $\bar{x}$'s *appear* to fall from μ, based on distortions created by using small-sample s's to estimate $\sigma_{\bar{x}}$. In this example, the middle 98% of the $\bar{x}$'s appear to fall in the interval between ±2.90 standard deviations of μ. In reality, the $\bar{x}$'s are not falling ±2.90 standard deviations from μ, but the underestimations of $\sigma_{\bar{x}}$ make them appear so. These distorted z scores are called t scores.

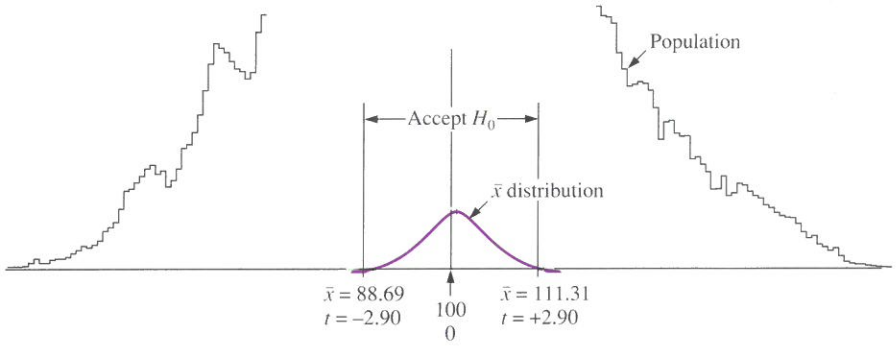
Rearranging the z formula, we have $z\sigma_{\bar{x}} = \bar{x} - \mu$. If $\sigma_{\bar{x}}$ is underestimated, z must increase to keep $\bar{x} - \mu$ constant. This might be written: $z\uparrow\sigma_{\bar{x}}\downarrow = \bar{x} - \mu$ (a constant value). Thus $z\uparrow = t$. Put thousands of these distorted z's into a distribution = t distribution. Each sample size has a separate t distribution.

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$-2.90 = \frac{\bar{x} - 100}{3.9} \qquad +2.90 = \frac{\bar{x} - 100}{3.9}$$

Solving for $\bar{x}$: $\bar{x} = 88.69$ $\bar{x} = 111.31$

The complete solution showing the cutoffs might appear as follows:



Sequence

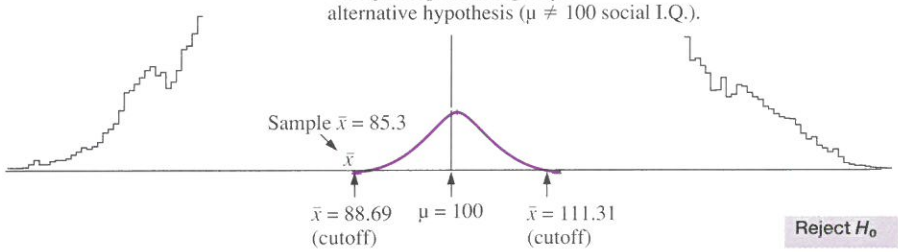
III. Accept or reject H_0 using your sample $\bar{x}$

METHOD ONE
This method uses the actual value of the sample $\bar{x}$ (85.3 social IQ) in the decision-making process.

Recall: our sample results were as follows:
n = 9 students
$\bar{x}$ = 85.3 social IQ

Criteria: Accept H_0 (μ = 100 social I.Q.) if your sample $\bar{x}$ falls between the established $\bar{x}$ cutoffs of 88.69 and 111.31, otherwise reject.

Decision: Since our sample $\bar{x}$ (85.3) fell in the rejection zone, we reject H_0 and accept H_1, the alternative hypothesis ($\mu \neq 100$ social I.Q.).

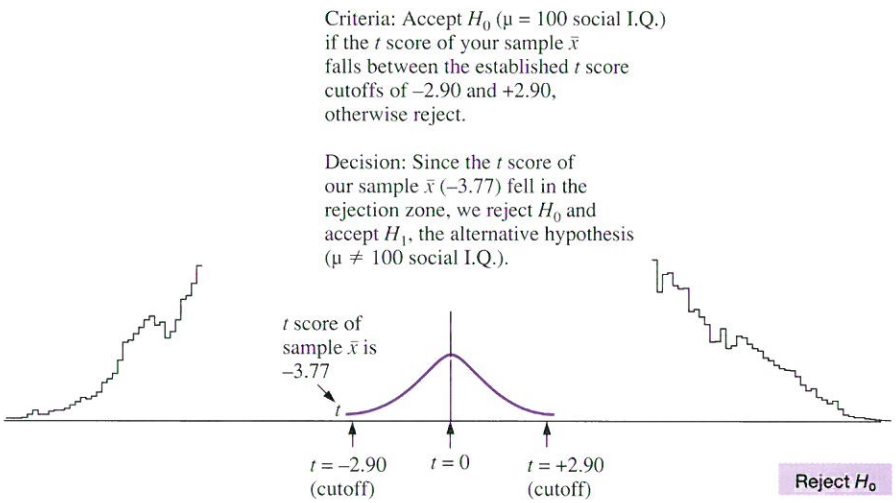


METHOD TWO

This method uses the t score of the sample $\bar{x}$ in the decision-making process. To use this method, however, we must first calculate the t score of our sample $\bar{x}$ (85.3 social IQ), as follows:

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{85.3 - 100}{3.9}$$

$$t = -3.77$$



Whether we use the actual value of the sample $\bar{x}$ or the equivalent t score, we will always make the same decision. In this case, we reject H_0. This implies we accept H_1: ($\mu \neq 100$).

Answer

a. The final answer may be presented using actual values or t scores.

Actual values — Since the sample average (sample $\bar{x}$ = 85.3 social IQ) was outside the range (88.31 to 113.31) where we would most likely expect sample averages to fall if H_0 were true, we reject H_0. Therefore, we must

Accept H_1: $\mu \neq 100$ social IQ

t scores — Since the t score of the sample average (−3.77) was outside the range (−2.90 to +2.90) where we would most likely expect t scores of sample averages to fall if H_0 were true, we reject H_0. Therefore, we must

Accept H_1: $\mu \neq 100$ social IQ

The answer may also be expressed simply as:

The null hypothesis is rejected,

or

$t = -3.77$ (significant)

So, to answer part a, yes, the results are significant.

b. If the data constitutes a valid random sample of underachievers, the results provide evidence that on average underachievers score significantly below other children in their age group in the social skills observed in the experiment, which we labeled social IQ.

However, like many experiments in the social sciences and education, this experiment is replete with potential violations to validity, as follows.

Random selection: First, the study violates our most basic tenet of sampling: random selection. Intact groups (such as, classes, clubs, and residents of a building) often have common interests or ties and cannot be assumed to represent a cross section of the target population. For instance, this particular underachiever class may be located in a high-family-income district and all nine students may be from privileged families. In other words, this one underachiever class may very well *not* represent the general population of underachievers.*

Unfortunately, a good many statistical studies in education, marketing, psychology, medicine, and other fields still employ this method of using intact groups, which is often why experiments in these fields that measure the same phenomenon vary so widely in results. Random selection must be assured if we are to use a sample as representative of a population. Intact groups do not constitute random selection (refer to chapter 1 for a more detailed discussion on random selection).

Other aspects of validity: Of course, even if random selection from the underachiever population can be assured, the potential for violations to other aspects of validity in such experiments is enormous. Essentially, we must guarantee internal and external validity, defined in regards to this experiment as follows:

> Internal validity: the certainty that our observations were *accurate* and *reliable* measures of the social characteristics we set out to measure.
>
> External validity: the certainty that our *methods and presence* in no way affected the true social behavior of the children.

Assuring internal and external validity in experiments of this nature is no easy task. Actually, this experiment and its inherent questions of validity open up the whole topic of problems besetting scientific investigation in the social sciences. Although much too broad a topic to address here, I will discuss it briefly, then recommend two books for further reference.

To begin with, let us say, ''labels'' are dangerous in any scientific study. In this experiment, we called a certain set of social characteristics, social IQ, when, in fact, we really do not know what we measured, except for a collection of social characteristics at a certain point in a child's life. In reality, what we *may* be measuring is merely adaptability to white middle-class behavior in America rather than a universal set of social intelligence that crosses all cultures and classes. And this social intelligence may be ''learned'' behavior rather than ''inborn,'' thus the phrase IQ, which implies a natural capacity, may be misleading. But, be that as it may, whatever we label these characteristics, they first must be precisely defined and set forth at the beginning of the experiment so future researchers (or any reader for that matter) can properly assess and criticize your work.

*Intact groups may more effectively be used as *one* element in a random sample, say for instance, if we randomly selected 36 classes nationwide, in which case, one class offers one result in a random sample of $n = 36$ classes.

Now, once the characteristics we wish to measure are set forth, we must determine the best way to get an *accurate and reliable* measure of these characteristics. When personal judgment of an observer is involved, this is a difficult task. For instance, let's use how close a child should stand while talking. What objective scale can be used? Is the observer in any way biased? Would a child be rated the same from observer to observer? Does a rating of 120 on this characteristic mean twice the social savvy as a rating of 60? And what does "twice" the social savvy mean? Obstacles abound in this type of experiment.

Even if accurate and reliable measures can be attained (thus, ensuring internal validity), did the *methods* of the experiment or *presence* of the observer in any way alter the true social behavior of the child? Were the children observed in secret in their natural environment? Or were artificial situations enacted? How were the times chosen to observe the child—convenience to the observer, when the child is at play, when the child is upset? These are all variables that may substantially influence results.

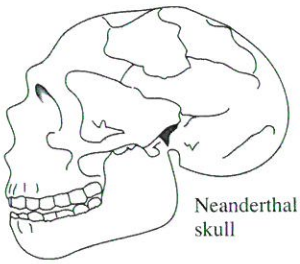
For further reading on proper experimental technique for testing in psychology, education, and other fields where similar difficulties exist in obtaining random selection and controlling risks to validity, refer to D. Campbell and J. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Boston: Houghton Mifflin Co., 1963) and L. Tyler and W. B. Walsh, *Tests and Measurements* (Englewood Cliffs, N.J.: Prentice-Hall, 1979). ■

Starting from Raw Data

Our second example concerns the cranial capacity (brain size) of Neanderthal man. The example is presented not only to demonstrate small-sample testing, but to demonstrate how a hypothesis test is performed starting from raw data.

Example

Much of the 1800s was spent by certain medical doctors and scientists trying to prove the cranial capacity of modern Caucasian man was larger (and, by implication, of greater intellect) than earlier Caucasian groups or other cultural groups, such as, Mongolian, Semitic, Malay, American, African, etc. Many educated people in the 1800s accepted this without evidence, but nothing was more illuminating than the fossil skulls first dug up in 1856 of Neanderthal man (a cousin to our Cro-Magnum species but much older and more primitive, who roamed Europe and the Middle East from 200,000 B.C. to 30,000 B.C.). Suppose 11 such Neanderthal skulls yielded the following:



Neanderthal skull

Neanderthal Skull Cranial Capacity (in cubic inches)	
85	91
89	94
93	90
90	88
91	86
	93

(assume a normal population)

a. Calculate $\bar{x}$ and s for this sample.

b. At a .05 level of significance, test the claim the sample came from a population with mean of 87.0 cubic inches *or less* (87.0 cubic inches is the average cranial capacity of modern Caucasian man). Are the results significant?

c. If the data constitutes a valid random sample from the Neanderthal population, what conclusions can be drawn? Briefly discuss validity.

Solution

a. We calculate $\bar{x}$ and s for this data, as follows:

Neanderthal Skull Cranial Capacity (in cubic inches)	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
85	90	−5	25
89	90	−1	1
93	90	3	9
90	90	0	0
91	90	1	1
91	90	1	1
94	90	4	16
90	90	0	0
88	90	−2	4
86	90	−4	16
93	90	3	9
$\Sigma x = 990$		$\Sigma(x - \bar{x})^2 = 82$	

$$\bar{x} = \frac{\Sigma x}{n} = \frac{990}{11}$$

$$\bar{x} = 90.0 \text{ cubic inches}$$

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

$$s = \sqrt{\frac{82}{11 - 1}} = \sqrt{8.2}$$

$$s = 2.864 \text{ cubic inches}$$

We summarize the results for part a as follows:

$n = 11$ Neanderthal skulls
$\bar{x} = 90.0$ cubic inches (cranial capacity)
$s = 2.86$ cubic inches

b. Since a small-sample size was used ($n = 11$), we must be careful that two conditions are first satisfied: (1) the population is at least somewhat mound shaped (it was stated, assume a normal population, so this condition is satisfied) and (2) since σ is not given and we must use s to estimate σ, we use t scores, not z scores.

A hypothesis test consists of three fundamental sequences as follows.

Sequence I. Set up initial conditions

State null hypothesis, the initial claim you wish to test:	H_0: $\mu = 87.0$ *or less* ($\mu \leq 87.0$)
State alternative hypothesis. If H_0 proves false, what must we conclude?	H_1: μ is more than 87.0 ($\mu > 87.0$)
State the risk of rejecting H_0 in error, the level of significance, α.	$\alpha = .05$ (5%)

Sequence

II. Assume H_0 true, use α to establish cutoff(s)

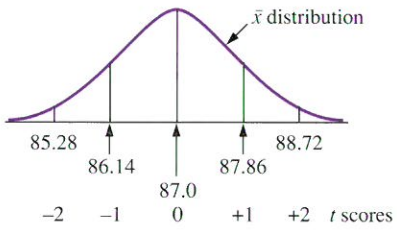
Calculate

$\sigma_{\bar{x}}$: $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} \approx \frac{2.864}{\sqrt{11}} \approx \frac{2.864}{3.317} \approx .863$$

Draw curves

Using our above calculation, $\sigma_{\bar{x}} \approx .86$, we estimate the spread of the $\bar{x}$ distribution.

(Keep in mind, $\sigma_{\bar{x}}$ is probably *under*estimated, however we will compensate for this by using t scores.)

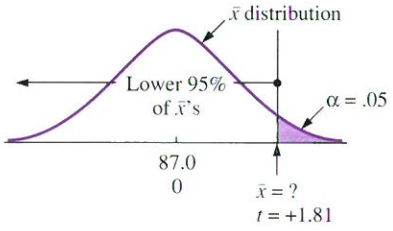


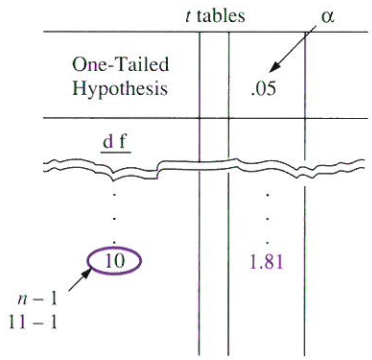
Since we are dealing with a one-tailed test (Note: $\mu = 87.0$ *or less*), we must assign the total α risk to one tail, in this case, to the right tail. Why the right tail? Because we reject H_0 only if our sample $\bar{x}$ falls significantly above 87.0 cubic inches.

Establish cutoffs using α

Our level of significance in this case is $\alpha = .05$ (5%), establishing the lower 95% of the $\bar{x}$'s as our region for accepting H_0 as true. In our t table, we look up a one-tailed hypothesis, $\alpha = .05$, to obtain $t = +1.81$. Remember: look down the df (degrees of freedom) column to 10 (df = $n - 1 = 11 - 1 = 10$).

Note: Had this been a large sample, the boundary would have been 1.65 standard deviations instead of 1.81.





We solve using the z formula, only now we use the letter t in place of z. We can view the t score as simply an adjustment to the z score (to compensate for the uncertainty created by using small-sample s's in the calculation of $\sigma_{\bar{x}}$).*

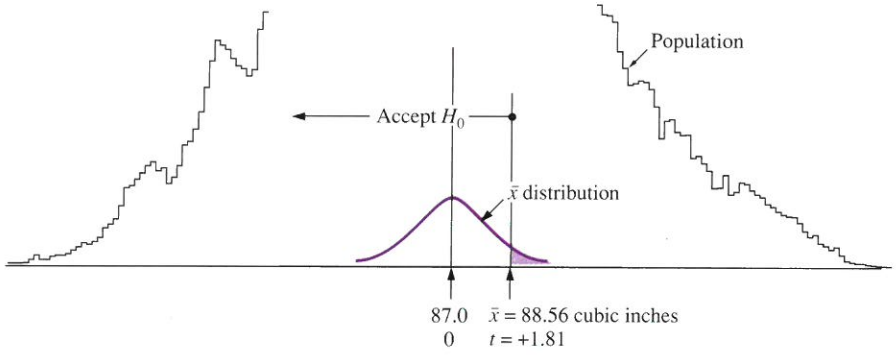
$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$+1.81 = \frac{\bar{x} - 87.0}{.863}$$

Note: (.863)(1.81) = 1.56 (rounded); adding this to 87.0 gives *88.56* cubic inches

Solving for $\bar{x}$: $\bar{x} = 88.56$

The completed solution might appear graphically as follows:



Sequence

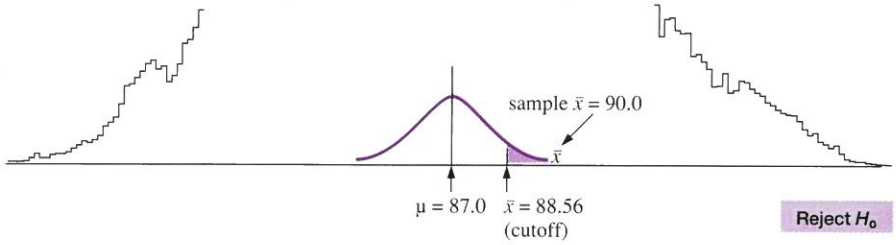
III. Accept or reject H_0 using your sample $\bar{x}$

METHOD ONE
This method uses the actual value of the sample $\bar{x}$ (90.0) in the decision-making process.

RECALL: our sample results were as follows:
n = 11 skulls
$\bar{x}$ = 90.0 cubic inches

Criteria: Accept H_0 (μ = 87.0 or less) if your sample $\bar{x}$ falls below the established cutoff of 88.56 cubic inches, otherwise reject.

Decision: Since our sample $\bar{x}$ (90.0) fell in the rejection zone, we reject H_0 and accept H_1, the alternative hypothesis (μ is more than 87.0 cubic inches).



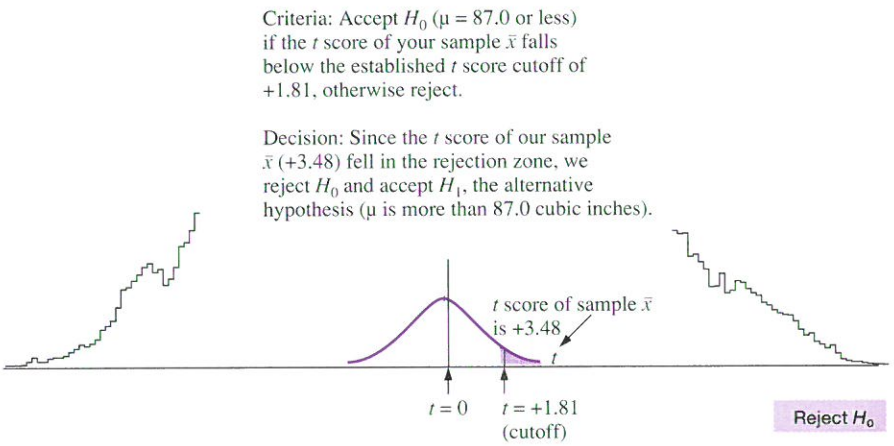
*In statistics, we view ourselves as sampling values from a precisely defined distribution. In this case, we are sampling from the t distribution (for n = 11), which is, in essence, a histogram of thousands and thousands of z scores describing how far in standard deviations the $\bar{x}$'s *appear* to fall from μ, based on distortions created by using small-sample s's to estimate $\sigma_{\bar{x}}$. In this example, the lower 95% of the $\bar{x}$'s appear to fall below +1.81 standard deviations from μ.

METHOD TWO

This method uses the t score of the sample $\bar{x}$ in the decision-making process. To use this method, however, we must first calculate the t score of our sample $\bar{x}$ (90.0 cubic inches), as follows:

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{90 - 87}{.863}$$

$$t = +3.48$$



Whether we use the actual value of the sample $\bar{x}$ or its equivalent t score, we will always make the same decision, in this case, we reject H_0. This implies we accept H_1 (μ is more than 87.0 cubic inches).

Answer

a. $\bar{x} = 90.0$; $s = 2.864$

b. The final answer may be presented using actual values or t scores.

Actual values — Since the sample average (90.0 cubic inches) was *above* the range where we would expect sample averages to fall if H_0 were true (up to 88.56 cubic inches), we reject H_0. Therefore, we must

Accept H_1: μ is more than an 87.0 cubic inch cranial capacity

t scores — Since the t score of the sample $\bar{x}$ (+3.48) was *above* the range where we would expect t scores to fall if H_0 were true (up to +1.81), we reject H_0. Therefore, we must

Accept H_1: μ is more than an 87.0 cubic inch cranial capacity

The answer may also be expressed simply as:

The null hypothesis is rejected.

or

$t = +3.48$ (significant)

So, to answer part b, yes, the results are significant.

c. If the data constitutes a valid random sample of Neanderthal skulls, the results indicate Neanderthal man had a larger brain size than modern Caucasian man. This, by the way, is true according to numerous skull

findings over the past century and a half. Cro-Magnum man (our direct ancestors), according to accumulated evidence, also had larger cranial capacity.

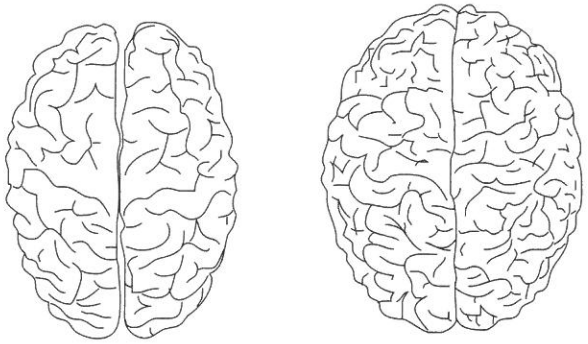
Validity: The primary issue here seems to be, again, random selection. If all eleven skulls came from the same fossil site, this does *not* constitute random selection. Coming from one fossil site, all the skulls may have been from members of one family or clan and it is not unusual for members of the same family or clan to have similar biological traits (e.g., large heads). In other words, validity would be more assured if the skulls were randomly selected from, let's say, a great many Neanderthal skulls discovered at several widely scattered sites.

Other risks to validity seem minimal. *Accurate and reliable* measures of cranial capacity can be achieved with properly calibrated scientific instruments, of course barring researcher mistakes, shoddy technique, or questions of honesty.

Historical discussion: This and other evidence led scientists in the late 1800s to conclude that brain size does not determine intelligence. Some other evidence was women on average have 5 to 10 cubic inches less cranial capacity than men and larger people tend to have larger brains. Perhaps one final deciding factor was the death of Carl Gauss in 1855 (universally acclaimed mathematician, scientist, and genius who, by the way, is responsible for the discovery and validation of much of the work you've studied in the last few chapters, specifically, the normal curve, standard deviation, and the central limit theorem). Upon autopsy, it was discovered Gauss's brain was near average in size. The theory that a large brain produces great intellect soon thereafter began to crumble. So, even in his death, the great Carl Gauss contributed to the advancement of knowledge. Although of near average brain size, he was a true giant among men.

One last word in Gauss's defense: Upon autopsy of his brain, it was noted, however, the surface of Gauss's brain was more richly textured than the average man's brain, with many more folds and crevices, as illustrated below.

Brain of a Papuan* (tribesman of the East Indies)



Brain of Carl Gauss

*From E. A. Spitzka, *Transactions of the American Philosophical Society* 21 (1907): 175–308, as presented in S. J. Gould, *The Mismeasurement of Man* (New York: W. W. Norton Co., 1981).

Perhaps a brain is like a radiator. A radiator with more folds and crevices (thus, more surface area) radiates more heat. Perhaps a brain with more folds and crevices radiates more intellect. For further reading on these topics of intelligence and cranial measurement, refer to S. J. Gould, *The Mismeasurement of Man*.

Summary

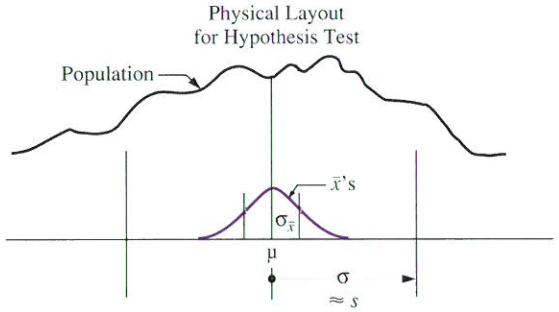
Hypothesis Test

A test designed to prove or disprove some initial claim. This initial claim is referred to as the null hypothesis and denoted as H_0.

We begin any hypothesis test by assuming the initial claim, the null hypothesis (H_0) is true. The second step is to establish a range of values where we would expect sample results (in this case, $\bar{x}$'s) to fall if H_0 were true. If the sample $\bar{x}$ of your experiment falls in this range, we merely accept H_0, otherwise we reject.

Physical Layout for Hypothesis Test (Large Sample, $n \geq 30$)

Both the population and $\bar{x}$ distributions have the same mean, μ, established by the null hypothesis, H_0.



The spread of the population, σ, can be estimated by s, the spread of the sample data. And the spread of the $\bar{x}$ distribution, $\sigma_{\bar{x}}$, is calculated using the central limit theorem formula,

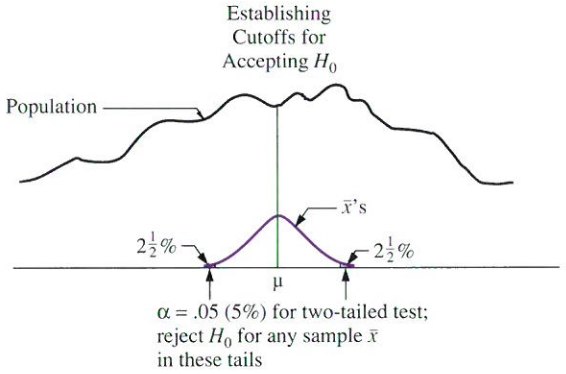
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

where s, the spread of the sample data, can be used to estimate σ, the spread of the data in the population.

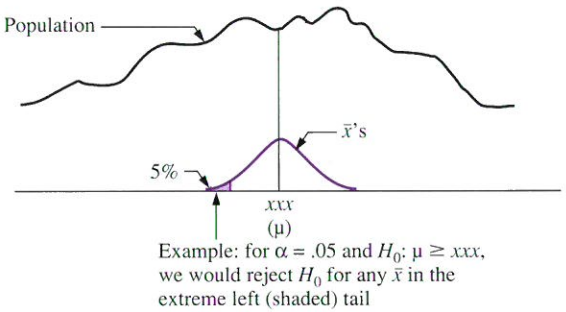
Establishing Cutoffs for Accepting H_0

The cutoffs are established using the level of significance (α risk) you are willing to accept in the experiment. For instance, if you establish $\alpha = .05$, you are willing to accept a 5% risk that when H_0 is true, you will reject it in error (refer to chapter 6 for a full discussion of errors).

For a two-tailed test, $\alpha = .05$ implies $2\frac{1}{2}\%$ of the risk is placed in each of the two extreme tails, establishing regions where you would reject H_0.



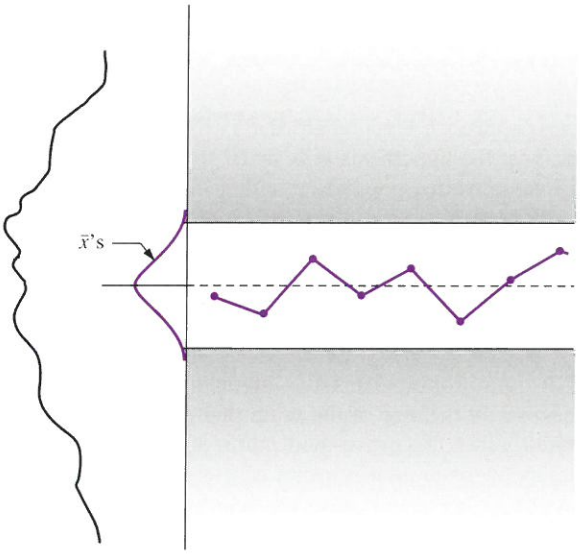
For a one-tailed test, the entire α risk is placed in one tail and we reject H_0 if our sample $\bar{x}$ falls in that one tail.



Determining which tail in which to place the α risk is dependent on how H_0 is stated. For example, if H_0 is stated, μ = *xxx or more,* this requires the entire α risk be placed in the extreme left tail (where you would reject H_0, shown in the previous sketch).

Control Charts

A control chart provides a clear visual history of a repetitive test. Essentially, cutoffs are established in a hypothesis test (using actual values or z scores) and the graph rotated $\frac{1}{4}$-turn counterclockwise, extending the cutoff lines to the right. This provides a clear on-going space to plot several sample $\bar{x}$'s. These $\bar{x}$'s, represented as dots in the sketch below, are often connected to each other by line segments as shown.

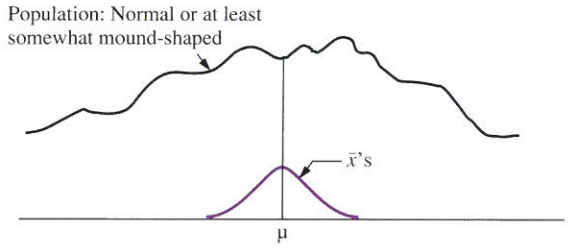


Small-Sample Hypothesis Testing ($n < 30$)

Small samples ($n < 30$) can be effectively used in hypothesis testing provided two conditions are satisfied, namely

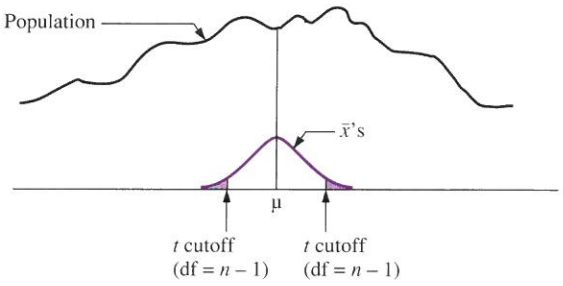
1. The population from which you sample is normally distributed or at least somewhat mound shaped. Generally, the smaller your sample size, the more critical this restriction, and
2. If we use s to estimate σ (which we almost always do) in the calculation of $\sigma_{\bar{x}}$, we must use a t score and not a z score to define the number of standard deviations the $\bar{x}$'s would be expected to fall from μ.

Discussion of condition 1: Although mild to moderate violation of this normal-population condition can often be tolerated with little effect on the validity of the test, severe departure (such as when a population is extremely skewed) can seriously compromise the test's validity.



Discussion of condition 2: Because of the tendency of small-sample s's to underestimate σ, we use a t score at the cutoffs (and not a z score), which essentially compensates for this tendency and allows us to maintain the validity level (α risk) of the test.

The t score values can be obtained from the t tables in back of the text (df = $n - 1$). Once these two conditions are satisfied, the small-sample hypothesis test is conducted in a manner much like any hypothesis test. For a two-tailed hypothesis test, the layout would be as follows.



Exercises

Note that full answers for exercises 1–5 and abbreviated answers for odd-numbered exercises thereafter are provided in the Answer Key.

7.1 A supplier claims the average thickness (diameter) of its fiber-optic thread is .560 mm (no more, no less), per your specifications. You receive a shipment but before you accept it, you take a random sample, with the following results:

n = 225 measurements
$\bar{x}$ = .553 mm
s = .030 mm

a. Test the supplier's claim at a .01 level of significance.
b. Are the results "significant"? Would you accept or reject the shipment?

7.2 Elavil is a powerful sedating drug prescribed by psychiatrists for the treatment of depression; however, dosages must be minimized to reduce side effects. A leading health organization claims the minimum effective dosage nationwide is on average 140.0 mg/day *or less*. A manufacturer of the drug decides to test the claim with the following sample results:

n = 900 patients
$\bar{x}$ = 141.6 mg/day minimum effective dosage
s = 48.2 mg/day

a. Test the health organization's claim ($\mu \leq 140.0$ mg/day) at $\alpha = .04$.
b. Are the results "significant"? Do we have evidence to refute the health organization's claim?

7.3
a. For exercise 7.1, set up a control chart.
b. On this control chart, plot the shipment result from exercise 7.1, $\bar{x}$ = .553 mm, along with additional shipment results: $\bar{x}$ = .558 mm, $\bar{x}$ = .551 mm, and $\bar{x}$ = .559 mm. Indicate the results that are significant.
c. For exercise 7.2, set up a control chart.
d. On this control chart, plot the sample results from exercise 7.2, $\bar{x}$ = 141.6 mg/day, along with the results of subsequently conducted studies: $\bar{x}$ = 138.7 mg/day, $\bar{x}$ = 137.1 mg/day, and $\bar{x}$ = 140.9 mg/day. Indicate the results that are significant.

7.4 Social IQ in children (the ability to read subtle, nonverbal clues and accurately assess social boundaries) may be a better predictor of success in life than mental IQ according to recent studies.

Suppose Mrs. Berry has her Raleigh-Durham elementary school class of 12 gifted (high mental IQ) students tested on a social IQ scale, with the following results:

n = 12 students
$\bar{x}$ = 110.8 social IQ
s = 16.3 (assume a normal population)

a. Test the hypothesis at $\alpha = .02$ that these students came from a population with social IQ, $\mu = 100$ (which is average for children of this age). Are the results significant?
b. If the data constitutes a valid random sample of gifted students, what conclusions can be drawn? Briefly discuss validity.

7.5 Anthropologists have long claimed it is not necesarily the size of the brain that determines intelligence. To prove their point, measurements are taken of the cranial capacity of n = 10 Neanderthal skulls (a species of primitive man living during the period from 200,000 B.C. to 30,000 B.C.) with the following results:

(In cubic inches)

87	98
88	87
95	84
91	95
89	96

(assume a normal population)

a. Calculate $\bar{x}$ and s for this sample.
b. At a .05 level of significance, test the claim that the sample group came from a population with a mean

of 87.0 cubic inches *or less* (87.0 cubic inches is the approximate average cranial capacity of modern man). Are the results significant?

c. If the data constitutes a valid random sample of Neanderthal skulls, what conclusions can be drawn? Briefly discuss validity.

7.6 Bloomindorfs Department Store in Manhattan hired a new specialist to redo their autumn fashion windows. The specialist is known for her exquisite coordinations of clothing, antique furniture, and accessories (all of which are sold at Bloomindorfs).

After setting up the new displays, Bloomindorfs wished to test whether any change had occurred in the number of customers entering the store per hour. Suppose it is known through electronic counters that the average number of customers for this store is 212 per hour and that a random sample now produces the following data.

n = 30 one-hour intervals
$\bar{x}$ = 231 customers per hour
s = 82 customers per hour

At a .05 level of significance, does the above data imply the new window displays have affected the average number of customers per hour entering the store? In other words, test the hypothesis, $\mu = 212$ per hour.

7.7 Excessive weight gain is a fear of those who wish to quit smoking, according to an article in the *New England Journal of Medicine*. Data gathered from the 1970s and 1980s indicate that after two years men gain on average 6.0 lb and women on average 8.0 lb more than those who continue to smoke.

Suppose a group of epidemiologists at the University of Rochester conducted the following research to see if these results still hold true in the 1990s.

Men: n = 110
$\bar{x}$ = 6.8 lb increase
s = 2.7 lb

Women: n = 90
$\bar{x}$ = 7.4 lb increase
s = 3.5 lb

a. At a .02 level of significance, test the claim that men still gain on average 6.0 lb.

b. At a .04 level of significance, test the claim that women still gain on average 8.0 lb.

c. Refer to part a. Suppose three such studies were conducted in the 1990s, yielding $\bar{x}$ = 6.8, $\bar{x}$ = 6.23, and $\bar{x}$ = 6.47. Set up a control chart demonstrating this.

d. Briefly discuss validity.

7.8 Pig farmers in Canada are trying to sell pork as the "other white meat besides chicken" (according to an article in *Western Producer*), claiming to have evolved a leaner, more efficient breed of pig. One of the claimed advantages of this new breed is that it takes substantially less time for a pig to reach final market weight (thus saving a farmer money in feeding and tending).

Suppose researchers at Texas Agriculture decided to test the claim, with the following results:

n = 80 randomly selected new-breed pigs
$\bar{x}$ = 157 days (birth to final market weight)
s = 11.5 days

The old breed took on average 170 days to reach final market weight. At a .03 level of significance, test the hypothesis that the average number of days to reach final weight for this new breed of pig is the same as the old breed, $\mu = 170$ days. (Notice we test the sample $\bar{x}$ against the established norm, in this case, that pigs currently take on average, $\mu = 170$ days to reach final market weight; it is common to use the established norm to set the hypothesis.)

7.9 Growth hormone administered to short children is thought to increase height beyond expected levels, according to published data (*Lancet* 336:1331–1334). Suppose research yielded the following results.

n = 38 children
$\bar{x}$ = 1.2 inch growth (beyond expected)
s = .84 inches

Generally in such experiments we assume a null hypothesis of "no change," in this case, 0" growth (beyond expected). In other words, it is customary in

scientific investigations to start with the assumption that the treatment is ineffective. In this case, that a child taking the hormone will grow, on average, no more or no less than expected for that time period. We must *prove* otherwise.

At $\alpha = .01$, test the claim that this sample was taken from a population of $\mu = 0''$ growth (beyond expected).

7.10 A toiletry manufacturer claims their bottles contain *at least* 9.00 oz. of bath lotion (as stamped on the label). Suppose the Federal Trade Commission (FTC) investigated by randomly sampling 49 bottles, with the following results:

$n = 49$ bottles
$\bar{x} = 8.94$ oz.
$s = .12$ oz.

At a .02 level of significance, does the FTC have legal grounds to proceed against the company on the unfair practice of short selling? In other words, test the claim, μ is at least 9.00 oz. ($\mu \geq 9.00$).

7.11 Certain nutritionists are outraged over the apparent lack of concern among physicians about so-called moderate cholesterol levels, claiming the average heart attack victim's cholesterol level is 230 mg/dl *or less*.

Suppose this prompted two studies, yielding the following results:

First Study: $n = 80$ heart attack victims
$\bar{x} = 226.0$ mg/dl cholesterol level prior to heart attack
$s = 37.3$ mg/dl

Second Study: $n = 67$ heart attack victims
$\bar{x} = 241.1$ mg/dl cholesterol level prior to heart attack
$s = 32.5$ mg/dl

a. Use the data in the first study to test the claim, $\mu \leq 230$ mg/dl ($\alpha = .01$).
b. Use the data in the second study to test the claim, $\mu \leq 230$ mg/dl ($\alpha = .05$).
c. What might account for the marked difference in results from the two studies?

7.12 Bypass surgery for obesity is increasingly becoming an option for those who have failed more moderate weight controlling strategies. In the procedure, vast tracts of the stomach are stapled off leaving a small pouch and narrow pathway leading to the intestines. Claims of weight losses have averaged over 120 lb (monitored two years after surgery).

Suppose researchers at the University of California tested this claim with the following results:

$n = 73$ patients monitored two years after surgery
$\bar{x} = 108$ lb weight loss
$s = 23$ lb

At $\alpha = .01$, test the claim, $\mu \geq 120.0$ lb weight loss.

7.13 To qualify for poverty funds, legislators in a particular district in Philadelphia had to show average household income for a family of four was $11,809 *or below*. A study was commissioned yielding the following:

$n = 53$ households in district
$\bar{x} = \$12{,}053$ annual household income
$s = \$4{,}320$

a. At $\alpha = .03$, does this particular legislative district qualify for poverty funds?
b. To reduce the possibility of falsely disqualifying a district, should a .03 or .01 level of significance be used? Briefly explain.

7.14 After the theft of several masterpieces, a New England Art Museum was quite concerned that the night security guard was diligent in performing his rounds. One way to monitor this was to ensure it took on average 21.0 minutes (no more, no less) to complete a known series of checks. The security guard was randomly clocked, yielding the following:

$n = 12$ observations
$\bar{x} = 18.6$ minutes
$s = 4.3$ minutes (assume a normal population)

a. At $\alpha = .01$, test the claim that the average time it takes for the security guard to make the rounds is 21.0 minutes ($\mu = 21.0$).

b. At $\alpha = .10$, test the same claim, $\mu = 21.0$.
c. To reduce the possibility of falsely accusing the guard of not adequately performing rounds, should a .01 or .10 α level be set? Briefly explain.

7.15 Jessica, a new recruit at a local army base in Georgia, claims her average time to clean her M16 rifle to pass inspection is "a cool 11.7 min." Her buddies, suspecting an unsubstantiated brag, decide to test her claim and randomly clocked her (in secret) on 6 attempts, yielding:

$n = 6$ rifle cleanings
$\bar{x} = 13.57$ minutes
$s = 3.2$ minutes (assume a normal population)

a. At $\alpha = .05$, test the claim that the average time it takes Jessica to clean her M16 rifle to pass inspection is at most 11.7 minutes ($\mu \leq 11.7$).
b. At $\alpha = .025$, test the same claim, $\mu \leq 11.7$.

7.16 A rare Assyrian coin minted well over 2000 years ago was claimed to have contained *at least* 3.2 grams of gold, on average. A museum curator managed to locate 8 such coins and assessed their gold content at:

2.9, 3.5, 3.0, 3.2, 3.3, 3.0, 2.7, 3.2 (assume a normal population)

a. At $\alpha = .01$ level of significance, test the claim $\mu \geq 3.2$.
b. If the data constitutes a valid random sample of these Assyrian coins, what conclusions can be drawn? Briefly discuss validity.

7.17 Children who are abused or severely neglected have lower intelligence and an increased risk of depression, drug abuse, and suicide, according to researchers at State University of New York/Albany and at University of Minnesota (*New York Times,* February 18, 1991, p. A11).

Suppose the American Association for the Advancement of Science, in an effort to substantiate these claims, sponsored a study in the Albany area, which monitored several randomly selected children. Of these, seven turned out to be abused or severely neglected, yielding the following:

Change in IQ: -6, -17, -12, -15, -9, -7, and -11 points
(assume a normal population)

a. In such experiments, we assume $\mu = 0$ change in IQ. In other words, it is customary in scientific investigations to start with the assumption the factors being studied (abuse or neglect) do not affect our measured variable (IQ). We must *prove* otherwise.
At $\alpha = .02$, test the hypothesis of $\mu = 0$ change in IQ.
b. If the data constitutes a valid random sample, what conclusions can be drawn? Briefly discuss validity.

7.18 Bone loss from prolonged space travel can be a serious problem, according to researchers at Pennsylvania State University, especially in the leg and spine regions. Suppose some estimate, on average, a 3% *or more* bone loss from extended space travel and the National Aeronautic and Space Administration (NASA) decided to conduct research on nine astronauts, yielding the following:

Bone Loss in Leg and Spinal Regions From Extended Space Travel (In Percentage Loss)

1.2	2.0	5.0
2.9	2.4	1.7
3.2	0.0	4.1

(assume a normal population)

a. At $\alpha = .05$, test the claim of $\mu = 3.0\%$ *or more* bone loss.
b. If the data constitutes a valid random sample, what conclusions can be drawn? Briefly discuss validity.

Endnotes

1. Historical endnote on W. Gossett: Gossett was part of a group of young university scientists in 1899 appointed to the brewing staff at the Guinness Breweries in Dublin. He tried to apply statistical techniques to experiments at the brewery and soon grew concerned using small-sample s's to estimate σ in experiments involving the quality of raw material (barley, hops, etc.), and in production tests and in finished-product tests.

This and other concerns led to a meeting with Karl Pearson in 1905. Gossett spent part of the academic year (1906–1907) at Pearson's Biometric Laboratory in London. From this stay, came a number of original papers, including the famous "Probable Error of a Mean" paper in 1908, in which he presented the t distributions for small-sample testing, which Fisher later incorporated into his ANOVA tables. We owe much to the original work of W. Gossett who published under the pen name, Student (assumedly referring to being a student of Karl Pearson).

Some early history: The t score distributions were derived empirically by W. Gossett and first published in an article, "The Probable Error of a Mean," which appeared in *Biometrika*, VI, pp. 1–25, in 1908, under the pen name, Student. Gossett was a British statistician and advisor to the Guinness Breweries in Dublin, Ireland. The Guinness Company forbade employees from publishing the results of research, however, the firm relaxed this ruling in Gossett's case to allow him to publish under a pen name.

Gossett derived the t distributions empirically using published data of body measurements (height and left middle finger) of 3000 criminals, from which he repeatedly selected small random samples. He successfully applied the results to other published data taken from (1) the *Journal of Physiology* (1904), which showed the effects of optical isomers of hyoscyamine hydrobromide in producing sleep, and data taken from (2) the *Journal of the Agricultural Society* concerning the causes which lead to the production of hard (glutinous) wheat and soft (starchy) wheat.

Gossett's article was a ground-breaking achievement. In many experiments, only small samples can be used, which Gossett cited as, "some chemical, many biological and most agricultural and large scale" experiments, and these had been outside the range of statistical enquiry, that is, up until his research.

See endnotes 1 and 10 in chapter 10 for more on W. Gossett.