UNIT – 3 T – TEST

❖ T – TEST FOR ONE SAMPLE :

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

o GAS MILEAGE INVESTIGATION :

- Federal law might eventually specify that new automobiles must average, for example, 45 miles per gallon (mpg) of gasoline.
- Because it's impossible to test all new cars, Compliance tests would be based on random samples from the entire production of each Car model.
- If a hypothesis test indicates substandard performance, the manufacturer Would be penalized, we'll assume, $200 per car for the entire production.
- In these tests, the null hypothesis states that, with respect to the mandated mean Of 45 mpg, nothing special is happening in the population for some car model—that Is, there is no substandard performance and the population mean equals or exceeds 45 mpg.

$$H_0 : \mu \geq 45$$
$$H_1 : \mu < 45$$

- The alternative hypothesis reflects a concern that the population mean is less than 45 mpg. Symbolically, the two statistical hypotheses reads :
- From the manufacturer's perspective, a type I error (a stiff penalty, even though the car Complies with the standard) is very serious.
- Accordingly, to control the type I error, Let's use the .01 instead of the customary .05 level of significance. From the federal Regulator's perspective, a type II error (not penalizing the manufacturer even though the car fails to comply with the standard) also is serious.
- In practice, a sample size Should be selected, to control the type II error, that is, to Ensure a reasonable detection rate for the smallest decline (judged to be important) of The true population mean below the mandated 45 mpg.
- To simplify computations in the Present example, however, the projected one-tailed test is based on data from a very Small sample of only six randomly selected cars.
- For this reasons that will become apparent, the z test must be replaced by a new hypothesis test, the t test.

❖ SAMPLING DISTRIBUTION OF T :

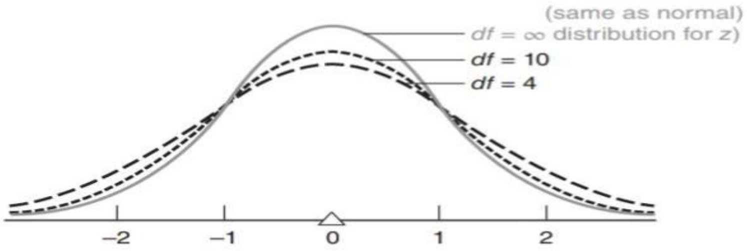
- Like the sampling distribution of z, the sampling distribution of t represents the distribution that would be obtained if a value of t were calculated for each sample mean for all possible random samples of a given size from some population.

- In the early 1900s, William Gosset discovered the sampling distribution of t and subsequently reported his Achievement under the pen name of “Student.” Actually, Gosset discovered not just one But an entire family of t sampling distributions (or “Student’s” distributions).
- Each t distribution is associated with a special number referred to as degrees of freedom.
- The concept of degrees of freedom is introduced because we’re Using variability in a sample to estimate the unknown variability in the population.
- when the n deviations about the sample mean are used to estimate variability In the population, only n – 1 are free to vary because of the restriction that the sum of these deviations must always equal zero.
- Since one degree of freedom is lost because of the zero-sum restriction, there are only n – 1 degrees of freedom, that is, symbolically, where df represents degrees of freedom and n equals the sample size.
- Since the gas mileage investigation involves six cars, the corresponding t test is based on a sampling distribution with five degrees of freedom (from df = 6 − 1).

DEGREES OF FREEDOM (ONE SAMPLE)

$$df = n - 1$$

- **COMPARED TO THE STANDARD NORMAL DISTRIBUTION :**



Various t distributions

- The above figure shows three t distributions. When there is an infinite (∞) number of Degrees of freedom (and, therefore, the sample standard deviation becomes the same as the population standard deviation), the distribution of t is the same as the standard normal distribution of z.
- Notice that even with only four or ten degrees of freedom, a t distribution shares a number of properties with the normal distribution.

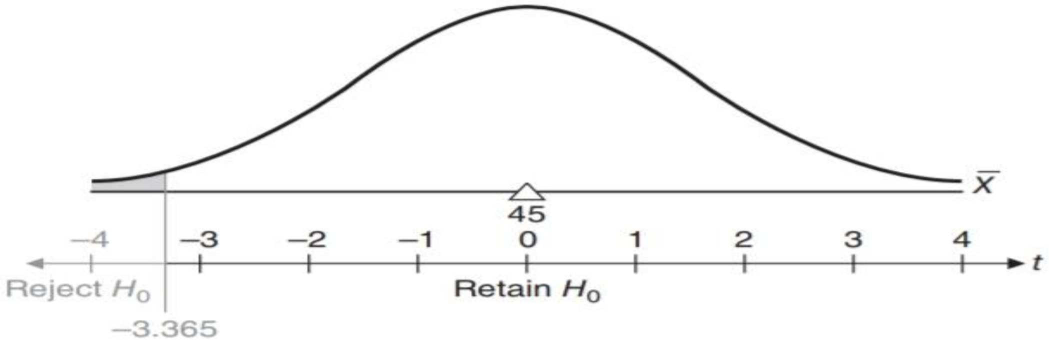
- All t distributions are symmetrical, unimodal, and bell-shaped, with a dense concentration that peaks in the middle (when t equals 0) and tapers off both to the right and left of the middle (as t Becomes more positive or negative, respectively).
- The inflated tails of the t distribution, Particularly apparent with small values of df, constitute the most important difference Between t and z distributions..

- **TABLE FOR T DISTRIBUTIONS :**

- To save space, tables for t distributions concentrate only on the critical values of T that correspond to the more common levels of significlevels
- The critical t values for either one- or two-tailed hypothesis tests at the .05, .01, and .001 levels of significance all listed critical t values are positive and originates from the upper half of each distribution. Because of the symmetry of the t distribution, we can obtain the corresponding critical t values for the lower half of each distribution merely by placing a negative sign in front of any entry in the table.

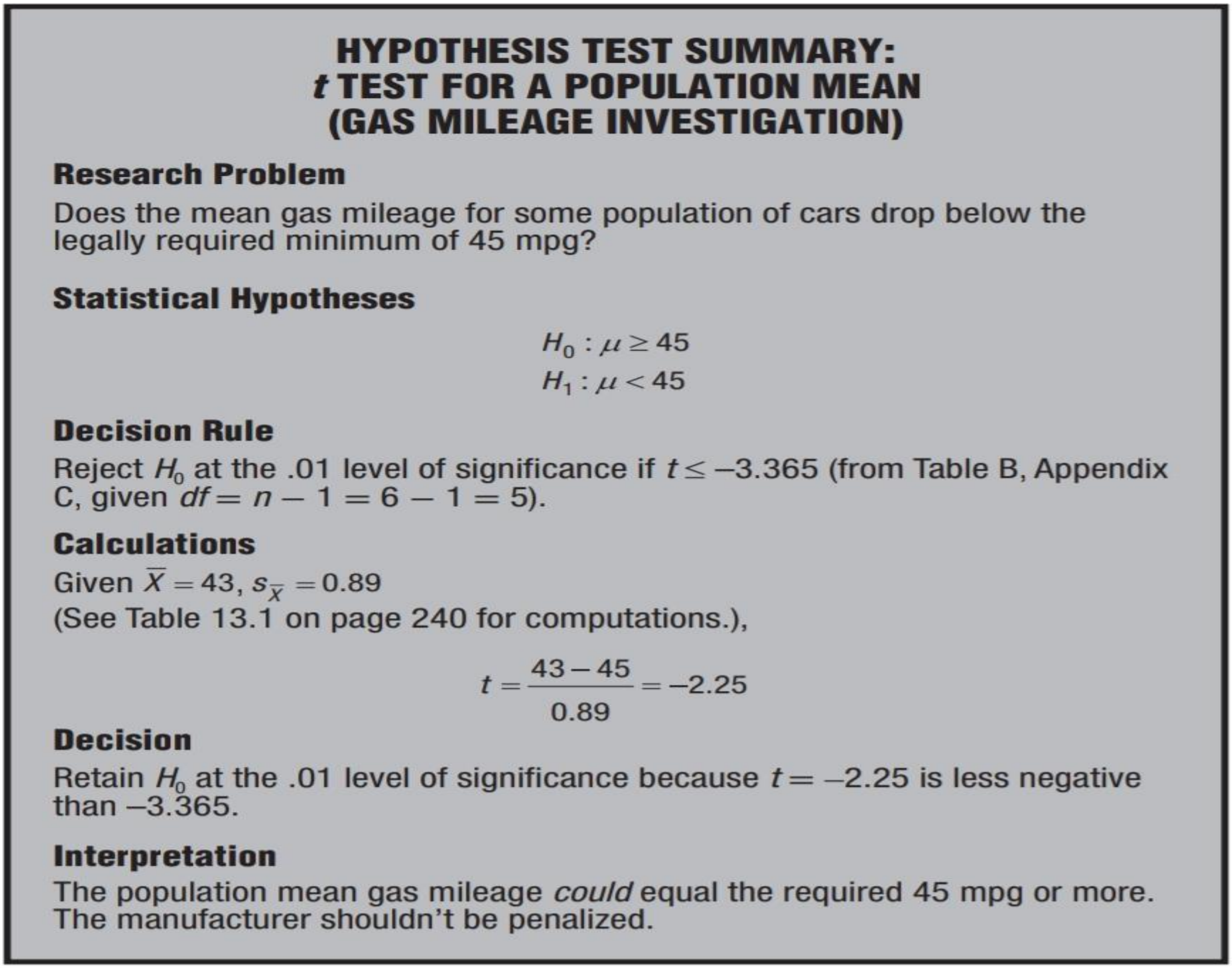
- **FINDING CRITICAL T VALUES :**

- To find a critical t value we have to read the entry in the cell intersected by the row for the correct number of degrees of freedom and the column for the test specifications.
- For example, to find the critical t for the gas mileage investigation, first go to the right-Hand panel for a one-tailed test, then locate both the row corresponding to five degrees Of freedom and the column for a one-tailed test at the .01 level of significance.
- The intersected cell specifies 3.365. A negative sign must be placed in front of 3.365, since the hypothesis test requires the lower tail to be critical.
- Thus, –3.365 is the critical t for the gas mileage investigation, and the corresponding decision rule is illustrated in the below figure, where the distribution of t is centered about zero (the equivalent value of T for the original null hypothesized value of 45 mpg).
- If the gas mileage investigation had involved a two-tailed test (still at the .01 level With five degrees of freedom), then the left-hand panel for a two-tailed test would have Been appropriate, and the intersected cell would have specified 4.032.
- Both positive And negative signs would have to be placed in front of 4.032, since both tails are critical. In this case, 4.032 would have been the pair of critical t values.



Hypothesized sampling distribution of t (gas mileage investigation)

- **MISSING df IN TABLE :**
  - If the desired number of degrees of freedom doesn't appear in the df column of Table, use the row in the table with the next smallest number of degrees of freedom.
  - For example, if 36 degrees of freedom are specified, use the information from the Row for 30 degrees of freedom. Always rounding off to the next smallest df produces a slightly larger critical t, making the null hypothesis slightly more difficult to reject.
  - This procedure defuses potential disputes about borderline decisions by investigators With a stake in rejecting the null hypothesis.



HYPOTHESIS TEST SUMMARY: *t* TEST FOR A POPULATION MEAN (GAS MILEAGE INVESTIGATION)

Research Problem

Does the mean gas mileage for some population of cars drop below the legally required minimum of 45 mpg?

Statistical Hypotheses

$$H_0 : \mu \geq 45$$
$$H_1 : \mu < 45$$

Decision Rule

Reject H_0 at the .01 level of significance if $t \leq -3.365$ (from Table B, Appendix C, given $df = n - 1 = 6 - 1 = 5$).

Calculations

Given $\overline{X} = 43$, $s_{\overline{X}} = 0.89$
(See Table 13.1 on page 240 for computations.),

$$t = \frac{43 - 45}{0.89} = -2.25$$

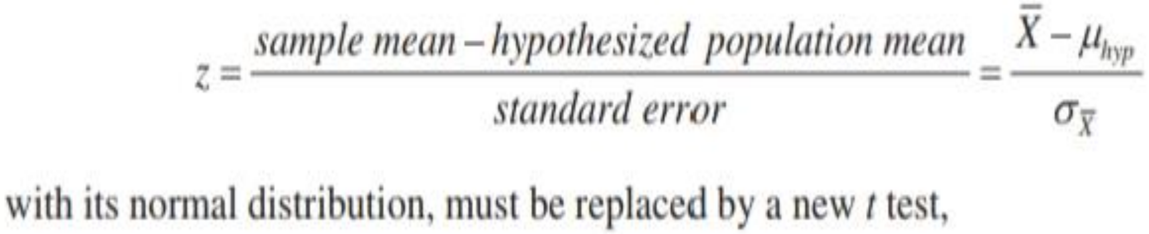
Decision

Retain H_0 at the .01 level of significance because $t = -2.25$ is less negative than -3.365.

Interpretation

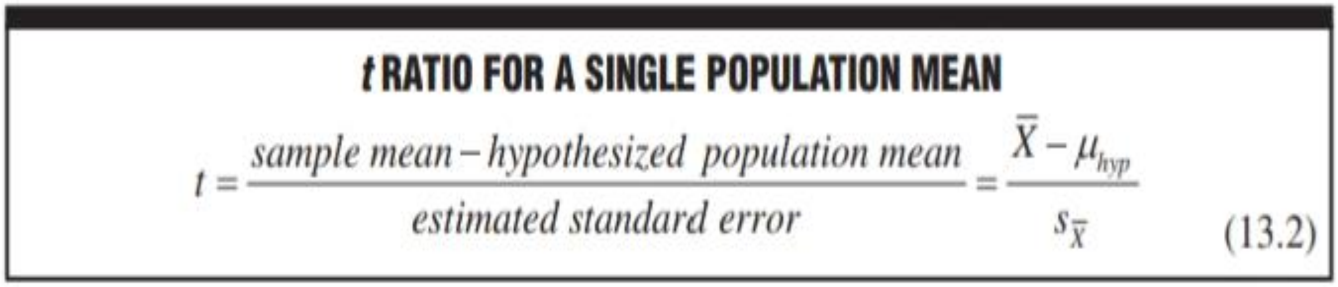
The population mean gas mileage *could* equal the required 45 mpg or more. The manufacturer shouldn't be penalized.

- **T – TEST :**
  - Usually, as in the gas mileage investigation, the population standard deviation is Unknown and must be estimated from the sample.
  - The subsequent shift from the standard error of the mean o its estimate, Xs , has an important effect on the entire Hypothesis test for a population mean. The familiar z test,



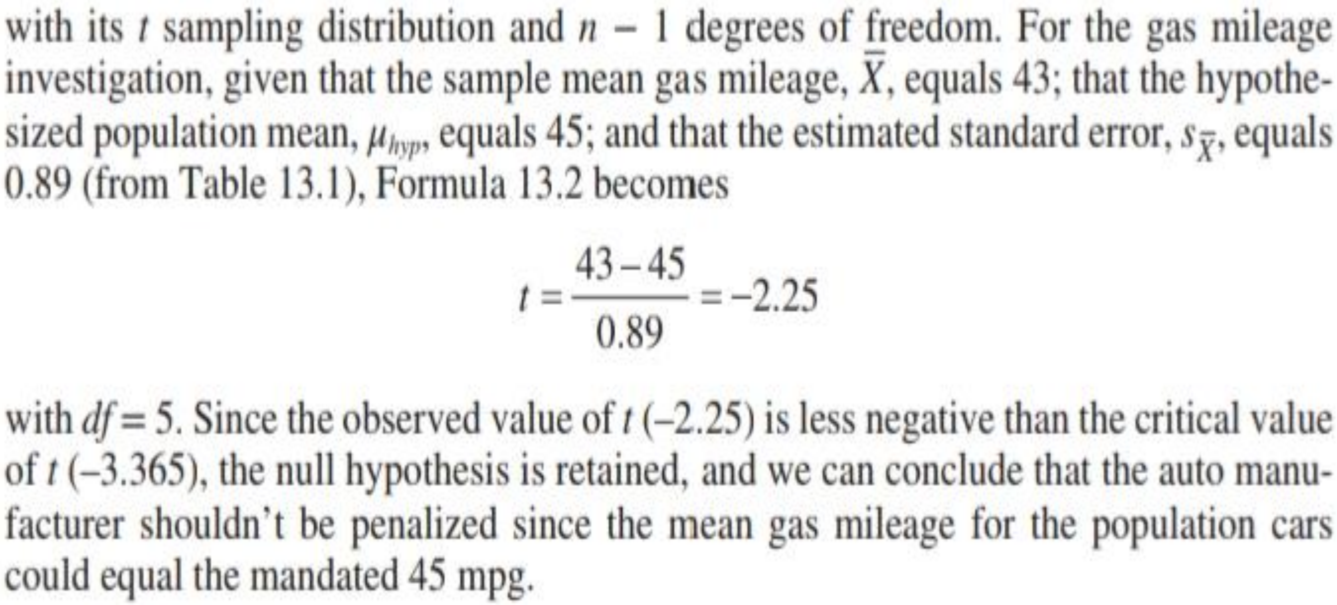
$$z = \frac{\textit{sample mean} - \textit{hypothesized population mean}}{\textit{standard error}} = \frac{\overline{X} - \mu_{hyp}}{\sigma_{\overline{X}}}$$

with its normal distribution, must be replaced by a new *t* test,



t RATIO FOR A SINGLE POPULATION MEAN

$$t = \frac{\textit{sample mean} - \textit{hypothesized population mean}}{\textit{estimated standard error}} = \frac{\overline{X} - \mu_{hyp}}{s_{\overline{X}}} \qquad (13.2)$$



with its *t* sampling distribution and *n* – 1 degrees of freedom. For the gas mileage investigation, given that the sample mean gas mileage, $\overline{X}$, equals 43; that the hypothesized population mean, μ_{hyp}, equals 45; and that the estimated standard error, $s_{\overline{X}}$, equals 0.89 (from Table 13.1), Formula 13.2 becomes

$$t = \frac{43 - 45}{0.89} = -2.25$$

with *df* = 5. Since the observed value of *t* (–2.25) is less negative than the critical value of *t* (–3.365), the null hypothesis is retained, and we can conclude that the auto manufacturer shouldn't be penalized since the mean gas mileage for the population cars could equal the mandated 45 mpg.

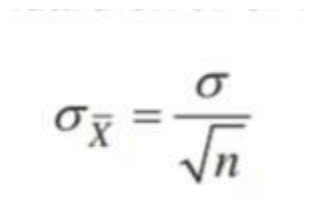
- **GREATER VARIABLILITY OF T RATIO :**
  - As has been noted, the tails of the sampling distribution for t are more inflated than those for z, particularly when the sample size is small.
  - Consequently, to accommodate the greater variability of t, the critical t value must be larger than the corresponding critical z value.
  - For example, given the one-tailed test at the .01 level of significance for the Gas mileage investigation, the critical value for t (–3.365) is larger than that for z (–2.33).
- **COMMON THEME OF HYPOTHESIS TEST :**
  - All of the hypothesis tests represent variations on the same theme, If some Observed characteristic, such as the mean for a random sample, qualifies as a Rare outcome under the null hypothesis, the hypothesis will be rejected. Other-Wise, the hypothesis will be retained.
  - To determine whether an outcome is rare, the observed characteristic is converted to Some new value, such as t, and compared with critical values from the appropriate sampling distribution.
  - Generally, if the observed value equals or exceeds a positive critical Value (or if it equals or is more negative than a negative critical value), the outcome Will be viewed as rare and the null hypothesis will be rejected.

❖ **DEGREE OF FREEDOM :**

- Typically, when it is used to estimate some unknown population characteristic, not all Observed values within the sample are free to vary.
- For example, the gas mileage data Consist of six values: 40, 44, 46, 41, 43, and 44. Nevertheless, the t test for these data has only five degrees of freedom because of the zero-sum restriction.
- Only five of these Six observed values are free to vary about their mean of 43 and, therefore, provide valid Information for purposes of estimation.
- The concept of degrees of freedom is introduced only because we are using observations in a sample to estimate some unknown Characteristic of the population and some Times several degrees of freedom will be lost.
- In any event, however, the degrees of Freedom always indicate the number of values free to vary, given one or more mathematical restrictions on a set of values used to estimate some unknown population Characteristic.

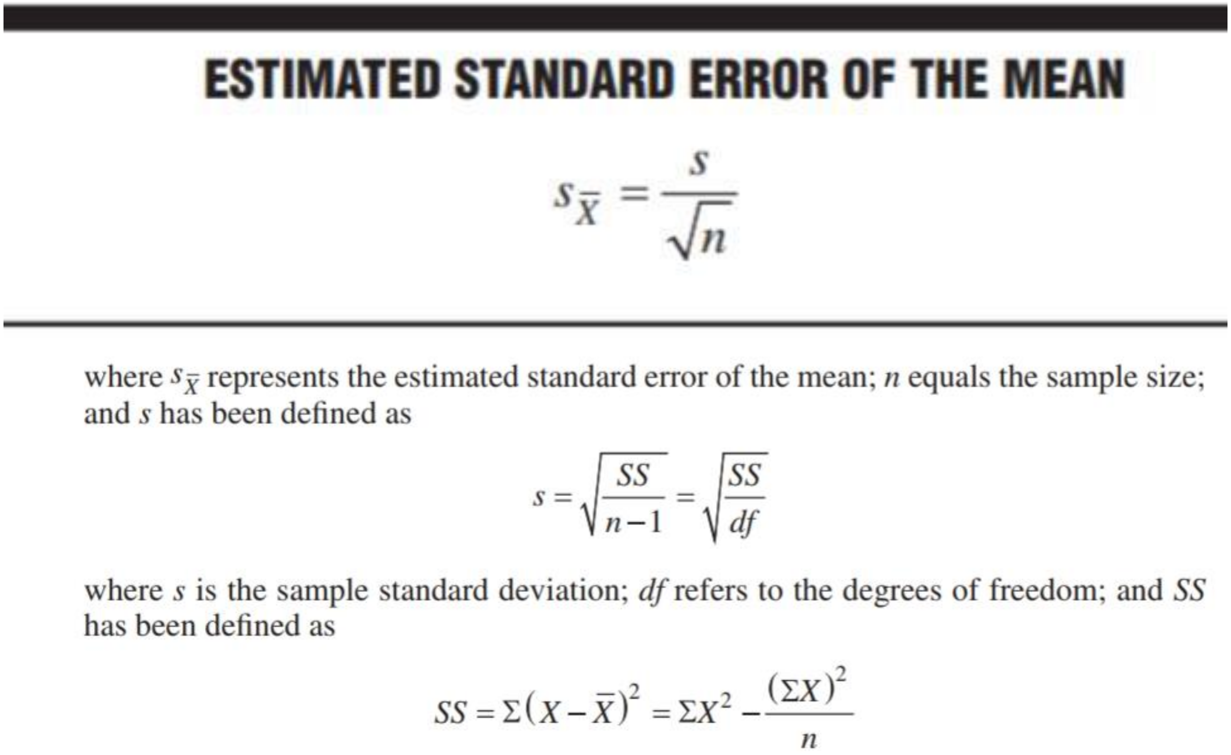
❖ **ESTIMATING THE STANDARD ERROR :**

- If the population standard deviation is unknown, it must be estimated from the sample.
- This seemingly minor complication has important implications for hypothesis testing indeed, it is the reason why the z test must be replaced by the t test.
- Now s Replaces σ in the formula for the standard error of the mean. Instead of,



$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

We have ,



ESTIMATED STANDARD ERROR OF THE MEAN

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

where $s_{\overline{X}}$ represents the estimated standard error of the mean; *n* equals the sample size; and *s* has been defined as

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

where *s* is the sample standard deviation; *df* refers to the degrees of freedom; and *SS* has been defined as

$$SS = \Sigma\left(X - \overline{X}\right)^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

- This new version of the standard error, the estimated standard error of the mean, is Used whenever the unknown population standard deviation must be estimated.
- **CALCULATIONS FOR THE T TEST :**

The three panels calculation for the t test shows the computational steps that produce a t of –2.25 for the gas mileage investigation.

PANEL I

This panel involves most of the computational labor, and it generates values for the Sample mean, X bar and the sample standard deviation, s. The sample standard deviation Is obtained to calculate the sum of squares, and after dividing the sum of squares, SS, by its degrees of freedom, n – 1, extracting The square root.

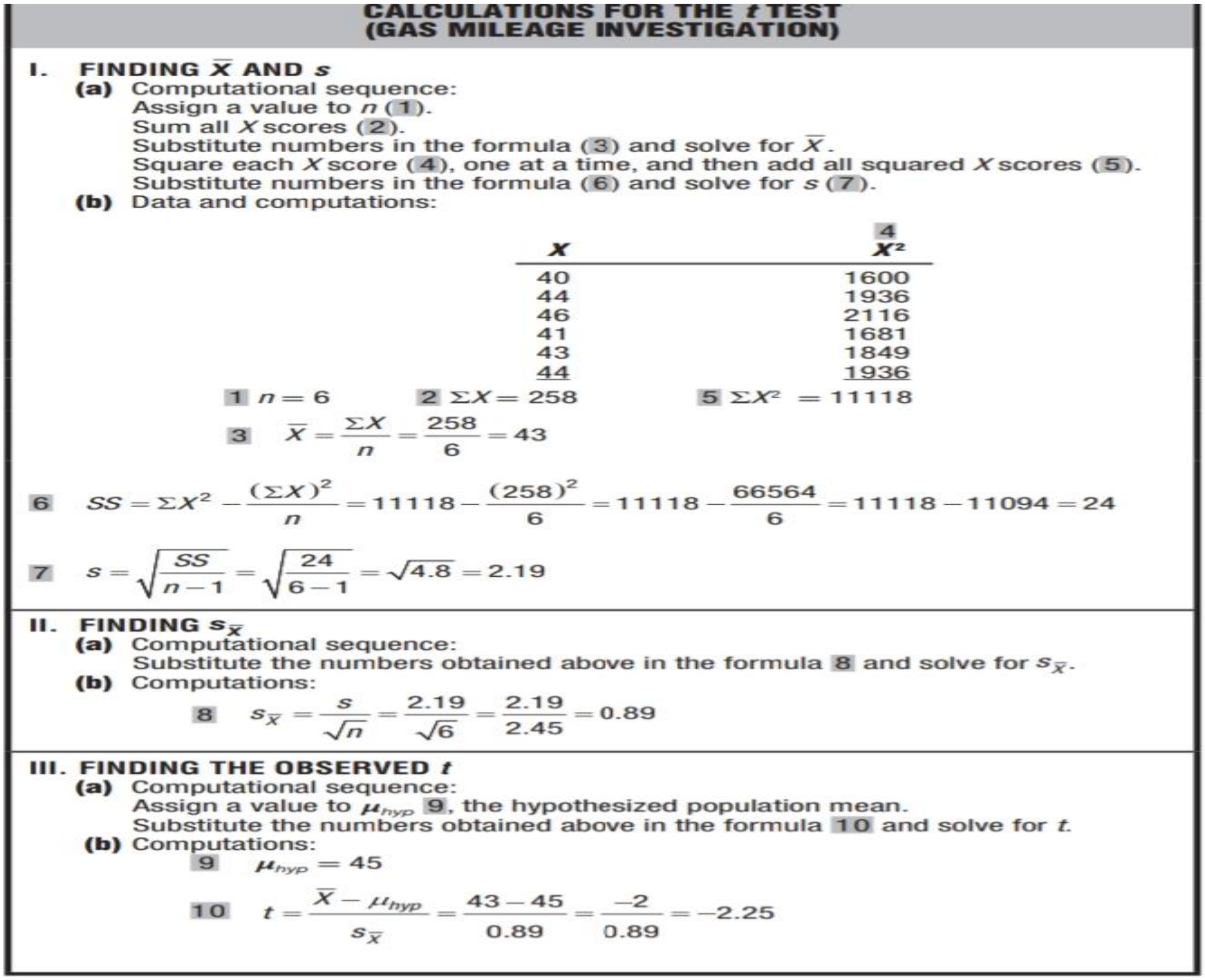
$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

PANEL II

Dividing the sample standard deviation, s, by the square root of the sample size, n, Gives the value for the estimated standard error, sx bar.

PANEL III

Finally, dividing the difference between the sample mean, X bar, and the null hypothesized value, $\boldsymbol{\mu}$hyp, by the estimated standard error, sx bar , yields the value of the t ratio.



CALCULATIONS FOR THE *t* TEST
(GAS MILEAGE INVESTIGATION)

I. FINDING $\overline{X}$ AND *s*

(a) Computational sequence:
Assign a value to n (1).
Sum all X scores (2).
Substitute numbers in the formula (3) and solve for $\overline{X}$.
Square each X score (4), one at a time, and then add all squared X scores (5).
Substitute numbers in the formula (6) and solve for s (7).

(b) Data and computations:

X	4 X^2
40	1600
44	1936
46	2116
41	1681
43	1849
44	1936

1 $n = 6$ 2 $\Sigma X = 258$ 5 $\Sigma X^2 = 11118$

3 $\overline{X} = \frac{\Sigma X}{n} = \frac{258}{6} = 43$

6 $SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 11118 - \frac{(258)^2}{6} = 11118 - \frac{66564}{6} = 11118 - 11094 = 24$

7 $s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{24}{6-1}} = \sqrt{4.8} = 2.19$

II. FINDING $s_{\overline{X}}$

(a) Computational sequence:
Substitute the numbers obtained above in the formula 8 and solve for $s_{\overline{X}}$.

(b) Computations:

8 $s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{2.19}{\sqrt{6}} = \frac{2.19}{2.45} = 0.89$

III. FINDING THE OBSERVED *t*

(a) Computational sequence:
Assign a value to μ_{hyp} 9, the hypothesized population mean.
Substitute the numbers obtained above in the formula 10 and solve for t.

(b) Computations:

9 $\mu_{hyp} = 45$

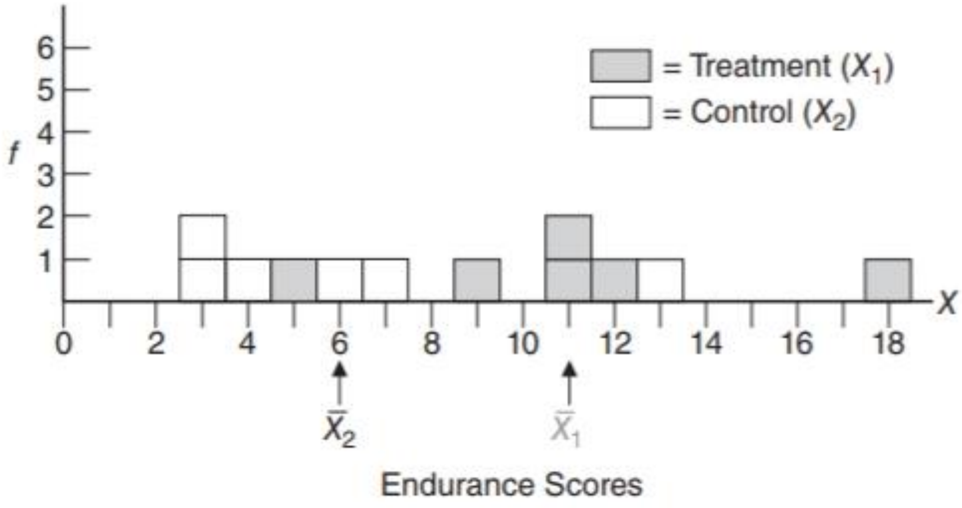
10 $t = \frac{\overline{X} - \mu_{hyp}}{s_{\overline{X}}} = \frac{43-45}{0.89} = \frac{-2}{0.89} = -2.25$

- **T TEST FOR TWO INDEPENDENT SAMPLES :**

- **EXPO EXPERIMENT :**

- Recent editions of the world's best-known bicycle race, the Tour de France, has seen Some cyclists expelled for attempting to enhance their performance by using a variety Of banned substances, including a synthetic "blood-doping" hormone, erythropoietin (EPO), that stimulates the production of oxygen-bearing (and fatigue-inhibiting) red Blood cells.
- Assume that a mental health investigator at a large clinic wants to determine whether EPO viewed as a potential therapeutic tool might increase the endurance of severely depressed patients.
- Volunteer patients are randomly assigned to one of Two groups: a treatment group (X1) that receives a prescribed amount of EPO and a control group (X2) that receives a harmless neutral substance.
- Subsequent endurance scores are based on the total time, in minutes, that each patient remains on a rapidly moving Treadmill.
- The statistical analysis focuses on the difference between mean endurance Scores for the treatment and control groups.
- For computational convenience, the results for the current experiment are based On very small samples of only six endurance scores per group (rather than on larger Sample sizes selected with the aid of power curves).
- Also For computational convenience, endurance scores have been rounded to the nearest Minute even though, in practice, they surely would reflect measurement that is more Precise.
- A glance at the below figure suggests considerable overlap in the scores for the two Groups.
- The treatment scores tend to be slightly larger than the control scores, and this Tendency is supported by the mean difference of 5 minutes (from 11 – 6) in favor of The treatment group.

- **TWO INDEPENDENT SAMPLES :**
  - In the current experiment, there are two independent samples, because each of the Two groups consists of different patients.
  - When samples are independent, observations In one sample are not paired, on a one-to-one basis, with observations in the other Sample.



❖ **STATISTICAL HYPOTHESIS :**

○ **NULL HYPOTHESIS :**

- ➢ According to the null hypothesis, nothing special is happening because EPO does Not facilitate endurance.
- ➢ In other words, either there is no difference between the means For the two populations (because EPO has no effect on endurance) or the difference Between population means is negative (because EPO hinders endurance).
- ➢ An equivalent statement in symbols reads:

$$H0 : \mu1 - \mu2 \leq 0$$

- ➢ Where H0 represents the null hypothesis and $\mu1$ and $\mu2$ represent the mean endurance Scores for the treatment and control populations, respectively.

○ **ALYERNATIVE (0R) RESEARCH HYPOTHESIS :**

- ➢ The investigator wants to reject the null hypothesis only if the treatment increases Endurance scores.
- ➢ Given this perspective, the alternative (or research) hypothesis Should specify that the difference between population means is positive because EPO Facilitates endurance.
- ➢ An equivalent statement in symbols reads:

$$H1 : \mu1 - \mu2 > 0$$

- ➢ Where H1 represents the alternative hypothesis and, as above, $\mu1$ and $\mu2$ represent the Mean endurance scores for the treatment and control populations, respectively.
- ➢ This Directional alternative hypothesis translates into a one-tailed test with the upper tail Critical.

○ **TWO OTHER POSSIBLE ALTERNATIVE HYPOTHESIS :**

1. Another directional hypothesis, expressed as

$$H1 : \mu1 - \mu2 < 0$$

Translates into a one-tailed test with the lower tail critical.

2. A nondirectional hypothesis, expressed as

$$H1 : \mu1 - \mu2 \neq 0$$

Translates into a two-tailed test.

- **SAMPLING DISTRIBUTION OF X1 BAR – X2 BAR :**
- Because of the inevitable variability associated with any difference between the sample mean endurance scores for the treatment and control groups, X1bar – X2bar, we can't Interpret a single observed mean difference at face value.
- The new mean difference for a repeat experiment would most likely differ from that for the original experiment.
- The sampling distribution of X1bar – X2bar is a concept introduced to account for the Variability associated with differences between sample means.
- It represents the entire Spectrum of differences between sample means based on all possible pairs of random samples from the two underlying populations.
- Once the sampling distribution has been centered about the value of the null hypothesis, we can determine whether the one observed sample mean difference qualifies as a common or a rare outcome. (A common outcome signifies that the observed sample mean difference could be due to variability or chance and, therefore, shouldn't be taken seriously. On the other hand, a rare outcome signifies that the observed sample mean difference probably reflects a Real difference and, therefore, should be taken seriously.)
- Since all the possible pairs Of random samples usually translate into a huge number of possibilities—often of Astronomical proportions—the sampling distribution of X1bar – X2bar isn't constructed from Scratch.
- As with the sampling distribution of XBar described in, statistical theory must be relied on for information about the mean and standard error for this new sampling distribution.
- **MEAN OF THE SAMPLING DISTRIBUTION, µX1bar – X2 bar :**
  - The mean of the sampling distribution of Xbar equals the Population mean, that is

    $$\mu \ xbar = \mu$$

    Where µxbar is the mean of the sampling distribution and µ is the population mean.
  - Similarly, the mean of the new sampling distribution of X1bar - X-bar equals the difference Between population means, that is,

    $$\mu \ Xbar - X2 \ bar = \mu1 - \mu2 ,$$

    Where X X 1 2M is the mean of the new sampling distribution and 1 2 µ µ is the difference Between population means. This conclusion is not particularly startling. Because of Sampling variability, it's unlikely that the one observed difference between sample Means equals the difference between population means. Instead, it's likely that, just By chance, the one observed difference is either larger or smaller than the difference Between population means. However, because not just one but all possible differences Between sample means contribute to the mean of the sampling distribution, X X 1 2 µ , the Effects of sampling variability are neutralized, and the mean of the sampling distribution Equals the difference between population means. Accordingly, these two terms are used Interchangeably. Any claims about the difference between population means, including The null hypothesized claim that this difference equals zero, can be transferred directly To the mean of the sampling distribution.

- **STANDARD ERROR OF THE SAMPLING DISTRIBUTION, σX1bar - X2bar :**
  - The standard deviation of the sampling distribution (or standard error) of X equals, Where x is the standard error, is the population standard deviation, and n is the Sample size.

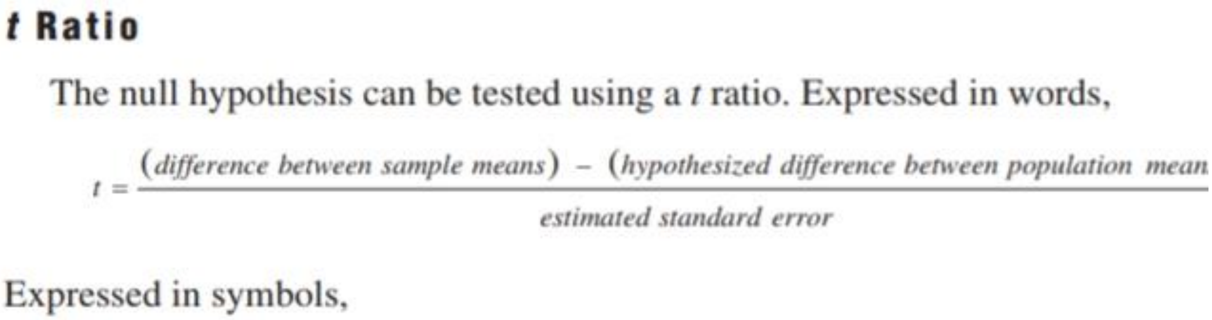
$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

  - To highlight the similarity between this expression and that for the new Sampling distribution, the population variance, 2, is introduced in the above equation by placing both the numerator and denominator under a common square root sign.
  - The standard deviation of the new sampling distribution of X1bar - x2bar equals where Sigma X1bar - x2bar is the new standard error, sigma1^2 and Sigma2^2 are the two population variances, and n1 and n2 are the two sample sizes.
  - The new standard error emerges directly from the original standard error with the Addition of a second term, sigma2^2 divided by n2, reflecting extra variability due to the shift From a single sample mean to differences between two sample means.
  - Therefore, the Value of the new standard error always will be larger than that of the original one.
  - The Original standard error reflects only the variability of single sample means about the Mean of their sampling distribution.
  - But the new standard error reflects extra variability When, as a result of random pairings, large differences between pairs of sample means Occur, just by chance, because they happen to deviate in opposite directions.
  - The standard error of the difference between Means, sigmax1bar – x2bar, as a rough measure of the average amount by which any sample mean Difference deviates from the difference between population means.
  - Viewed in this fashion, if the observed difference between sample means is smaller than the standard error, It qualifies as a common outcome well within the average expected by chance, and the null hypothesis, H0, is retained.
  - On the other hand, if the observed difference is sufficiently larger than the standard error, it qualifies as a rare outcome well beyond the Average expected by chance, and H0 is rejected size of the standard error for two samples, sigmax1bar – x2bar much like that of the standard error for one sample described earlier—becomes smaller with increases in sample Sizes.
  - With larger sample sizes, the values of X1bar – x2bar tend to cluster closer to the difference between population means, µ1 – µ2 following more precise generalizations.

- **T TEST :**

- The hypothesis test for the current experiment will be based not on the sampling distribution of X1bar– X2bar but on its standardized counterpart, the sampling distribution of t there also is a sampling distribution of z, its use requires that both population deviations be known.

- Since, in practice, this information is rarely available, The z test is hardly ever appropriate, and only the t test will be described.

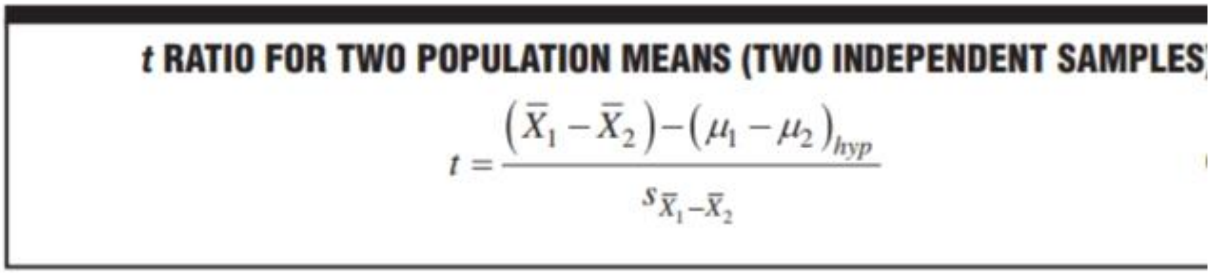


t Ratio

The null hypothesis can be tested using a *t* ratio. Expressed in words,

$$t = \frac{(\textit{difference between sample means}) - (\textit{hypothesized difference between population mean}}{\textit{estimated standard error}}$$

Expressed in symbols,



t RATIO FOR TWO POPULATION MEANS (TWO INDEPENDENT SAMPLES

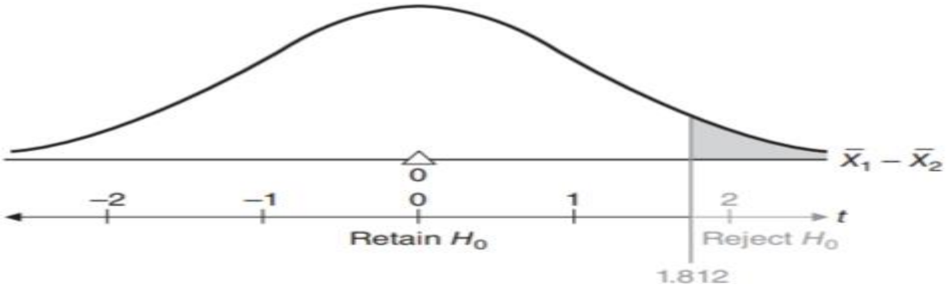
$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_{hyp}}{s_{\overline{X}_1 - \overline{X}_2}}$$

Which complies with a t sampling distribution having degrees of freedom equal to The sum of the two sample sizes minus two, that is, df = n1 + n2 – 2.

- The above formula X1bar - x2bar represents the one observed difference between sample means;(μ1 – μ2) hyp represents the hypothesized difference of Zero between population means; and s X1bar – x2bar represents the estimated standard error.

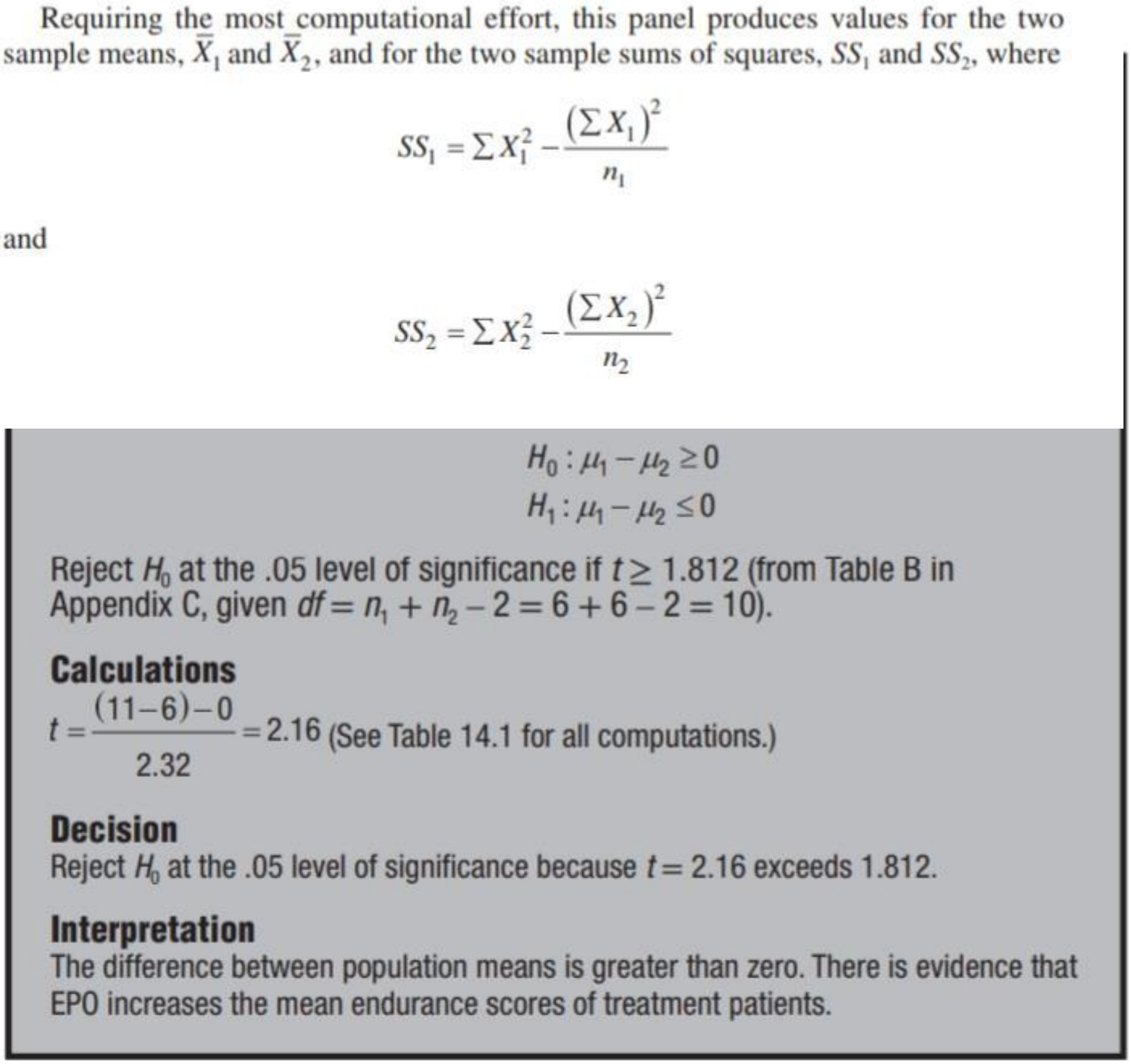
○ **FINDING CRITICAL T VALUES :**

- To determine critical values That distinguish between common and rare values of t on the assumption that the null Hypothesis is true.
- Read the entry in the cell intersected by the row for the correct number of Degrees of freedom, adjusted for two independent samples, and the column for the test Specifications.
- To find the critical t for the current experiment, first go to the right-hand Panel for a one-tailed test; next, locate the row corresponding to 10 degrees of freedom (from df = n1 + n2 – 2 = 6 + 6 – 2 = 10); and then locate the column for a one-tailed Test at the .05 level of significance.
- The intersected cell specifies 1.812. The corresponding decision rule is illustrated in below figure, where the sampling distribution of T is centered about the null hypothesized value of zero.
- The calculated t of 2.16 would Exceed the critical t of 1.812. Therefore, we can reject H0 and conclude that, on average, EPO increases the endurance scores of treatment patients.



❖ **CALCULATIONS FOR THE T TEST :**

PANEL I



Requiring the most computational effort, this panel produces values for the two sample means, $\bar{X}_1$ and $\bar{X}_2$, and for the two sample sums of squares, SS_1 and SS_2, where

$$SS_1 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}$$

and

$$SS_2 = \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}$$

$$H_0 : \mu_1 - \mu_2 \geq 0$$
$$H_1 : \mu_1 - \mu_2 \leq 0$$

Reject H_0 at the .05 level of significance if $t \geq 1.812$ (from Table B in Appendix C, given $df = n_1 + n_2 - 2 = 6 + 6 - 2 = 10$).

Calculations

$t = \dfrac{(11-6)-0}{2.32} = 2.16$ (See Table 14.1 for all computations.)

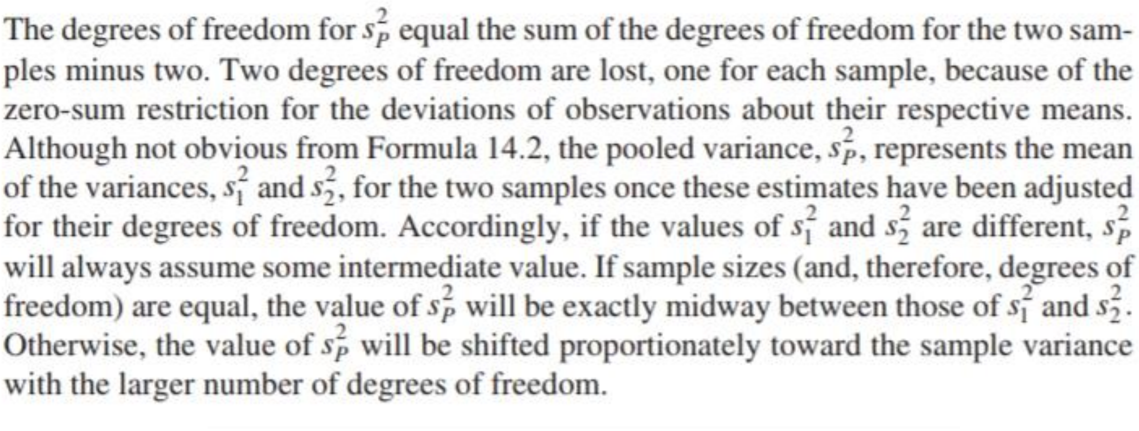
Decision

Reject H_0 at the .05 level of significance because $t = 2.16$ exceeds 1.812.

Interpretation

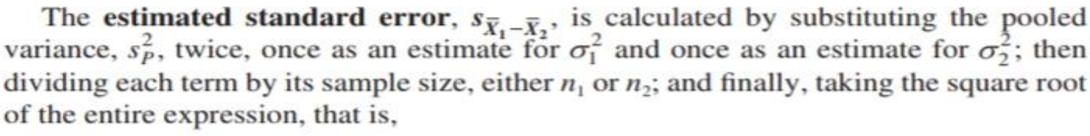
The difference between population means is greater than zero. There is evidence that EPO increases the mean endurance scores of treatment patients.

PANEL II

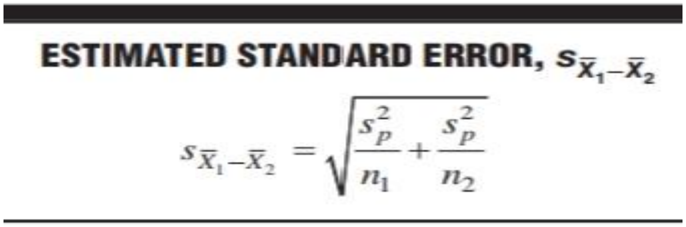


The degrees of freedom for s_p^2 equal the sum of the degrees of freedom for the two samples minus two. Two degrees of freedom are lost, one for each sample, because of the zero-sum restriction for the deviations of observations about their respective means. Although not obvious from Formula 14.2, the pooled variance, s_p^2, represents the mean of the variances, s_1^2 and s_2^2, for the two samples once these estimates have been adjusted for their degrees of freedom. Accordingly, if the values of s_1^2 and s_2^2 are different, s_p^2 will always assume some intermediate value. If sample sizes (and, therefore, degrees of freedom) are equal, the value of s_p^2 will be exactly midway between those of s_1^2 and s_2^2. Otherwise, the value of s_p^2 will be shifted proportionately toward the sample variance with the larger number of degrees of freedom.

PANEL III



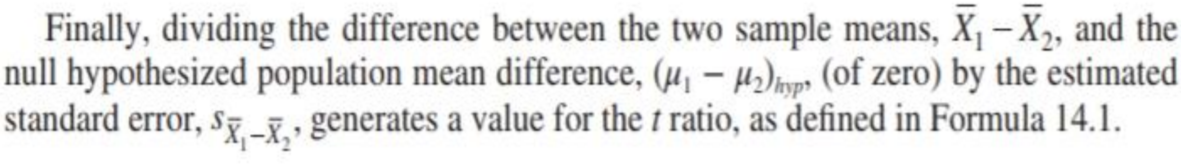
The **estimated standard error**, $s_{\overline{X}_1-\overline{X}_2}$, is calculated by substituting the pooled variance, s_p^2, twice, once as an estimate for σ_1^2 and once as an estimate for σ_2^2; then dividing each term by its sample size, either n_1 or n_2; and finally, taking the square root of the entire expression, that is,



ESTIMATED STANDARD ERROR, $s_{\overline{X}_1-\overline{X}_2}$

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

PANEL IV



Finally, dividing the difference between the two sample means, $\overline{X}_1 - \overline{X}_2$, and the null hypothesized population mean difference, $(\mu_1 - \mu_2)_{hyp}$, (of zero) by the estimated standard error, $s_{\overline{X}_1-\overline{X}_2}$, generates a value for the t ratio, as defined in Formula 14.1.

Table 14.1
CALCULATIONS FOR THE *t* TEST: TWO INDEPENDENT SAMPLES (EPO EXPERIMENT)

I. FINDING SAMPLE MEANS AND SUMS OF SQUARES, $\overline{X}_1$, $\overline{X}_2$, SS_1, AND SS_2

(a) Computational sequence:

Assign a value to n_1 (1).
Sum all X_1 scores (2).
Substitute numbers in the formula (3) and solve for $\overline{X}_1$.
Square each X_1 score (4), one at a time, and then add all squared X_1 scores (5).
Substitute numbers in the formula (6) and solve for SS_1.
Repeat this entire computational sequence for n_2 and X_2 and solve for $\overline{X}_2$ and SS_2.

(b) Data and computations:

ENDURANCE SCORES (MINUTES)

EPO	4	CONTROL	
X_1	X_1^2	X_2	X_2^2
12	144	7	49
5	25	3	9
11	121	4	16
11	121	6	36
9	81	3	9
18	324	13	169

1 $n_1 = 6$ 2 $\sum X_1 = 66$ 5 $\sum X_1^2 = 816$ $n_2 = 6$ $\sum X_2 = 36$ $\sum X_2^2 = 288$

3 $$\overline{X}_1 = \frac{\sum X_1}{n_1} = \frac{66}{6} = 11 \qquad \overline{X}_2 = \frac{\sum X_2}{n_2} = \frac{36}{6} = 6$$

6 $$SS_1 = \sum X_1^2 - \frac{(\sum X_1)^2}{n_1} = 816 - \frac{(66)^2}{6} = 816 - 726 = 90$$

$$SS_2 = \sum X_2^2 - \frac{(\sum X_2)^2}{n_2} = 288 - \frac{(36)^2}{6} = 288 - 216 = 72$$

II. FINDING THE POOLED VARIANCE, s_p^2

(a) Computational sequence:

Substitute numbers obtained above in the Formula (7) and solve for s_p^2.

(b) Computations:

7 $$s_p^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = \frac{90 + 72}{6 + 6 - 2} = \frac{162}{10} = 16.2$$

(*continued*)

III. FINDING THE STANDARD ERROR, $s_{\overline{X}_1 - \overline{X}_2}$

(a) Computational sequence:

Substitute numbers obtained above in the formula (8) and solve for $s_{\overline{X}_1 - \overline{X}_2}$.

(b) Computations:

8 $$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{16.2}{6} + \frac{16.2}{6}} = \sqrt{\frac{32.4}{6}} = \sqrt{5.4} = 2.32$$

IV. FINDING THE OBSERVED *t* RATIO

(a) Computational sequence:

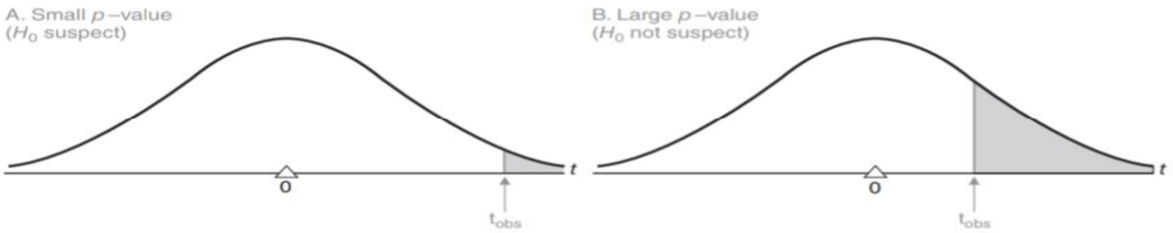
Substitute numbers obtained above in the formula (9), as well as a value of 0 for the expression $(\mu_2 - \mu_2)_{hyp}$ and solve for *t*.

(b) Computations:

9 $$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_{hyp}}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{(11 - 6) - 0}{2.32} = \frac{5}{2.32} = 2.16$$

❖ P - VALUES :

- The p-value for a test result represents the degree of rarity of that result, given That the null hypothesis is true. Smaller p-values tend to discredit the null Hypothesis and to support the research hypothesis.
- Most investigators adopt a less-structured approach to hypothesis testing, The null hypothesis is neither retained nor rejected, but viewed with degrees of suspicion, depending on the degree of rarity of the observed value of t or, more generally, the test result.
- Instead of subscribing to a single predetermined level of significance, the investigator waits until after the test result has been observed and then assigns a probability, known as a p-value, representing the degree of rarity attained by the test result.
- The p-value indicates the degree of rarity of the observed test Result when combined with all potentially more deviant test results.
- In other words, the P-value represents the proportion of area, beyond the observed result, in the tail of the Sampling distribution, the below figure by the shaded sectors for two different test results.
- In the left panel, a relatively deviant (from zero) observed it is associated with a small p-value that makes the null hypothesis suspect, while in the right panel, a relatively non-deviant observed t is associated with a large p-value that does not make the null hypothesis suspect.
- Below figure also Illustrates one-tailed p-values that are appropriate whenever the investigator has an interest only in deviations in a particular direction, as with a one-tailed Hypothesis test.
- Otherwise, two-tailed p-values are appropriate. Although not shown in Figure, two-tailed p-values would require equivalent shaded areas to be located in both tails of the sampling distribution, and the resulting two-tailed p-value would be Twice as large as its corresponding one-tailed p-value.



Shaded region showing small and large p vectors

❖ FINDING APPROXIMATE P VALUES :

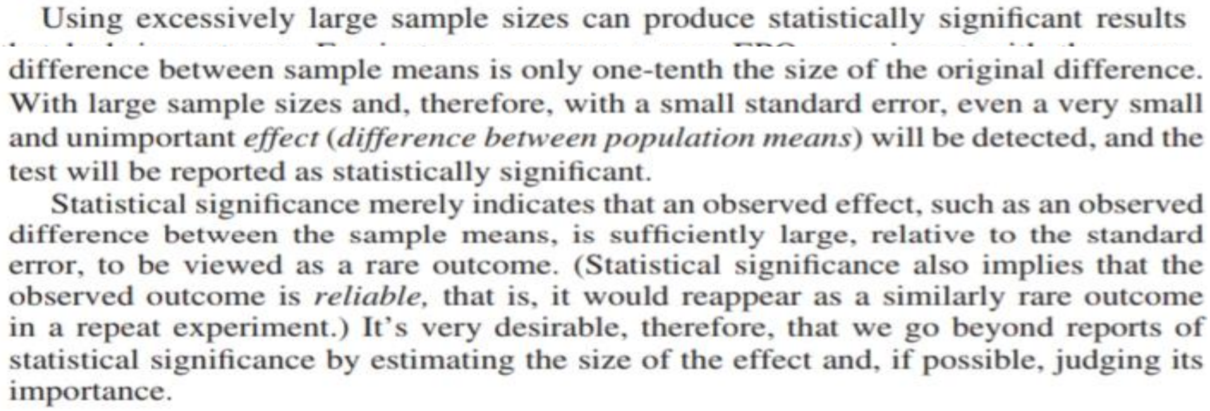
- To find approximate p-values, that is, p-values Involving an inequality, such as p < .05 or p > .05. To aid in the identification of these Approximate p-values, a shaded outline has been superimposed over the entries for t
- Once you've located the observed t relative to the tabular entries, simply follow the vertical line upward to identify the correct approximate p-value
- . find the approximate p-value for the t of 2.16 for the EPO experiment, first Identify the row in Table for a one-tailed test with 10 degrees of freedom.
- The three Entries in this row, 1.812, 2.764, and 4.144, serve as benchmarks for degrees of rarity to p-values of .05, .01, and .001, respectively.

- Since the observed t of 2.16 is located between the first entry of 1.812 and the second entry of 2.764, follow the Vertical line between the two entries upward to p < .05.
- From most perspectives, this Is a small p-value: The test result is rare—it could have occurred just by chance with a probability less than .05, given that H0 is true.
- Therefore, support has been mustered For the research hypothesis. This conclusion is consistent with the decision to reject H0 When a more structured hypothesis test at the .05 level of significance was conducted For the same data.

❖ **STATISTICALLY SIGNIFICANT RESULTS :**

- Implies only that the null hypothesis is probably false, and not Whether it's false because of a Large or small difference between Population means.
- It's important that you accurately interpret the findings of others—often reported as "having statistical significance."
- Tests of hypotheses often are referred to as tests of Significance, and test results are described as being statistically significant (if the null Hypothesis has been rejected) or as not being statistically significant (if the null hypothesis has been retained).
- Rejecting the null hypothesis and statistically significant both Signify that the test result can't be attributed to chance.
- However, correct usage dictates That rejecting the null hypothesis always refers to the population, such as rejecting the hypothesized zero difference between two population means, while statistically Significant always refers to the sample, such as assigning statistical significance to the difference between two sample means.
- Either phrase can be used. However, Assigning statistical significance to a population mean difference would be misleading, Since a population mean difference equals a fixed value controlled by "nature," not Something controlled by the results of a statistical test.
- Rejecting a sample mean difference also would be misleading, since a sample mean difference is an observed result that serves as the basis for statistical tests, not something to be rejected. Statistical significance doesn't imply that the underlying effect is important.
- Statistical significance between pairs of sample means implies only that the null hypothesis Is probably false, and not whether it's false because of a large or small difference Between population means.

○ **BEWARE OF EXCLUSIVELY LARGE SAMPLE SIZES :**

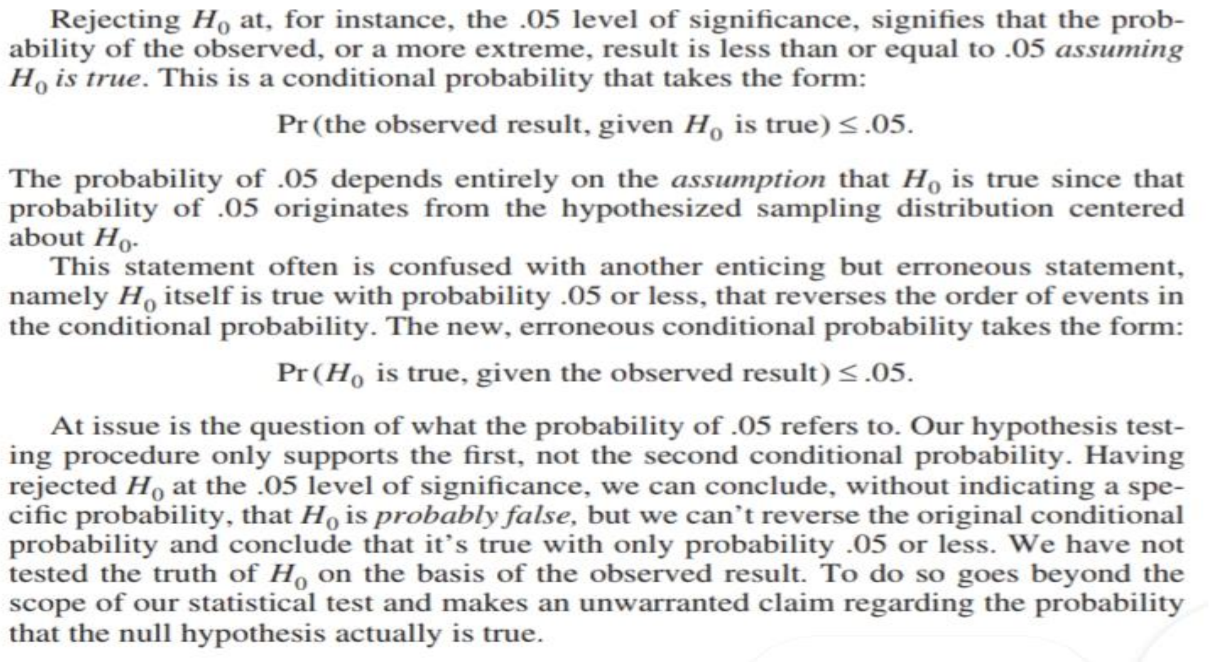


Using excessively large sample sizes can produce statistically significant results

difference between sample means is only one-tenth the size of the original difference. With large sample sizes and, therefore, with a small standard error, even a very small and unimportant *effect* (*difference between population means*) will be detected, and the test will be reported as statistically significant.

Statistical significance merely indicates that an observed effect, such as an observed difference between the sample means, is sufficiently large, relative to the standard error, to be viewed as a rare outcome. (Statistical significance also implies that the observed outcome is *reliable*, that is, it would reappear as a similarly rare outcome in a repeat experiment.) It's very desirable, therefore, that we go beyond reports of statistical significance by estimating the size of the effect and, if possible, judging its importance.

- **AVOID AN ERRONEOUS PROBABILITY CONDITION :**



Rejecting H_0 at, for instance, the .05 level of significance, signifies that the probability of the observed, or a more extreme, result is less than or equal to .05 *assuming H_0 is true*. This is a conditional probability that takes the form:

$$\Pr(\text{the observed result, given } H_0 \text{ is true}) \leq .05.$$

The probability of .05 depends entirely on the *assumption* that H_0 is true since that probability of .05 originates from the hypothesized sampling distribution centered about H_0.

This statement often is confused with another enticing but erroneous statement, namely H_0 itself is true with probability .05 or less, that reverses the order of events in the conditional probability. The new, erroneous conditional probability takes the form:

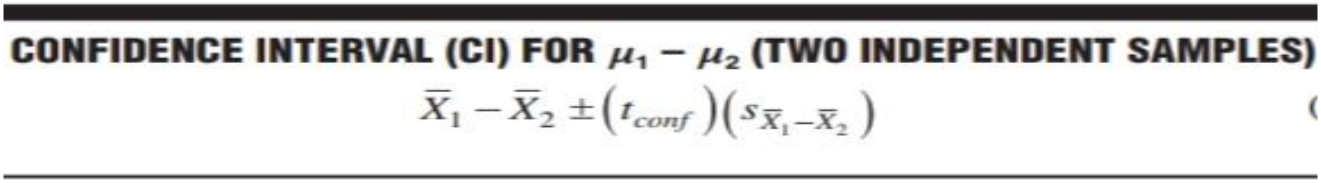
$$\Pr(H_0 \text{ is true, given the observed result}) \leq .05.$$

At issue is the question of what the probability of .05 refers to. Our hypothesis testing procedure only supports the first, not the second conditional probability. Having rejected H_0 at the .05 level of significance, we can conclude, without indicating a specific probability, that H_0 is *probably false*, but we can't reverse the original conditional probability and conclude that it's true with only probability .05 or less. We have not tested the truth of H_0 on the basis of the observed result. To do so goes beyond the scope of our statistical test and makes an unwarranted claim regarding the probability that the null hypothesis actually is true.

❖ **ESTIMATING EFFECT SIZE :**

- **POINT ESTIMATES (X1bar – X2bar) :**
  - a point estimate is the most straightforward type of estimate. It identifies the observed difference for X X 1 2, in this case, 5 minutes, as an estimate of the unknown effect, that is, the unknown difference Between population means, µ1 – µ2. On average, the treatment patients stay on the Treadmill for 11 minutes, which is almost twice as long as the 6 minutes for the control Patients.
  - this impressive estimate of effect size isn't surprising. With the very small groups of only 6 patients, we had to create a large, fictitious mean difference of 5 minutes in order to claim a statistically significant result. If this result had occurred in a real experiment, it would have signified a powerful effect of EPO on endurance that could be detected even with very small samples.
- **CONFIDENCE INTERVAL :**
  - Although simple, straightforward, and precise, point estimates tend to be inaccurate Because they ignore sampling variability.
  - Confidence intervals do not because,, they are based on the variability in the sampling distribution of X1bar – X2bar.
  - To estimate the range of possible effects of EPO on endurance, a confidence interval Can be constructed for the difference between population means, µ1 – µ2.
  - Confidence intervals for $\boldsymbol{\mu}1 - \boldsymbol{\mu}2$ specify ranges of values that, in the long run, Include the unknown effect (difference between population means) a certain Percent of the time.

- Given two independent samples, a confidence interval for µ1 – µ2 can be constructed From the following expression:



CONFIDENCE INTERVAL (CI) FOR $\mu_1 - \mu_2$ (TWO INDEPENDENT SAMPLES)

$$\bar{X}_1 - \bar{X}_2 \pm \left(t_{conf}\right)\left(s_{\bar{X}_1-\bar{X}_2}\right)$$

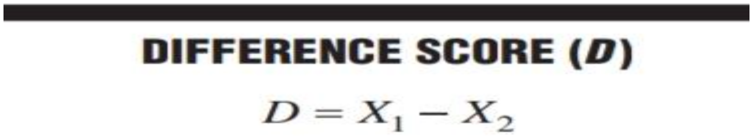
Where X1bar - x2bar represents the difference between sample means; tconf represents a number, distributed with n1 + n2 – 2 degrees of freedom, from the t tables, which satisfies the confidence specifications; and s(X1bar – X2bar) represents the estimated standard error.

- **INTEREPTING CONFIDENCE INTERVAL FOR $\boldsymbol{\mu}$1 – $\boldsymbol{\mu}$2 :**
  - The numbers in this confidence interval refer to differences between population Means, and the signs are particularly important since they indicate the direction of These differences.
  - Otherwise, the interpretation of a confidence interval for µ1 – µ2 is The same as that for µ. In the long run, 95 percent of all confidence intervals, similar To the one just stated, will include the unknown difference between population means.
  - Although we never really know whether this particular confidence interval is true or False, we can be reasonably confident that the true effect (or true difference between Population means) is neither less than –0.17 minutes nor more than 10.17 minutes.
  - If Only positive differences had appeared in this confidence interval, a single interpretation would have been possible.
  - However, the appearance of a negative difference in The lower limit indicates that EPO might hinder endurance, and therefore, no single Interpretation is possible.
  - Furthermore, the automatic inclusion of a zero difference in An interval with dissimilar signs indicates that EPO may have had no effect whatsoever on endurance.
  - The range of possible differences (from a low of –0.17 minute to a high of 10.17 Minutes) is very large and imprecise—as you would expect, given the very small sample sizes and, therefore, the relatively large standard error.
  - A repeat experiment should Use larger sample sizes in order to produce a narrower, more precise confidence interval that would reduce the range of possible population mean differences and effect Sizes.

- **META – ANALYSIS :**
  - A set of data-collecting and Statistical procedures designed To summarize the various effects Reported by groups of similar Studies.
  - The most recent Publication Manual of the American Psychological Association recommends that reports of statistical significance tests include some estimate of effect Size.
  - Because of the inevitable variability, attributable to differences in design, subject Populations, measurements, etc., as well as chance, the size of effects differs among Similar studies.

- Traditional literature reviews attempt to make sense out of these differences on the basis of expert judgment.
- Within the last couple of decades, literature Reviews have been supplemented by more systematic reviews, referred to as “meta-Analysis.”
- A meta-analysis begins with an intensive review of all relevant studies.
- This Includes small and even unpublished studies, to try to limit potential “publication bias” Arising from only reporting statistically significant results.
- Typically, extensive details Are recorded for each study, such as estimates of effect, design (for example, experi-Mental versus observational), subject population, variability, sample size, etc.
- Then the Collection of previous findings are combined using statistical procedures to obtain Either a composite estimate (for example, a standardized mean difference, such as Cohen’s d) of the overall effect and its confidence interval, or estimates of subsets of Similar effects, if required by the excessive variability among the original effects.

- **T TEST FOR TWO RELATED SAMPLES :**
  - **EXPO EXPERIMENT WITH REPEATED MEASURES :**
    - In the EPO experiment, the endurance scores of patients reflect not only the effect of EPO, if it exists, but also the random effects of many uncontrolled factors.
    - One very important type of uncontrolled factor, referred to as individual differences, Reflects the array of characteristics, such as differences in attitude, physical fitness, Personality, etc., that distinguishes one person from another.
    - If uncontrolled, individual differences can cause appreciable random variations among endurance scores and, Therefore, make it more difficult to detect any treatment effect. When each subject is Measured twice, the t test for repeated Measures can be extra sensitive to detecting a treatment effect by eliminating the distorting effect of variability due to individual differences.
- **DIFFERENCE SCORES (D) :**
  - Computations can be simplified by working directly with the difference between Pairs of endurance scores, that is, by working directly with where D is the difference score and X1 And X2 are the paired endurance scores for each Patient measured twice, once under the treatment condition and once under the control Condition, respectively.
  - Essentially, the use of difference scores converts a two-sample Problem with X1 and X2 Scores into a one-sample problem with D scores.
- **MEAN DIFFERENCE SCORE (DBAR) :**



DIFFERENCE SCORE (*D*)

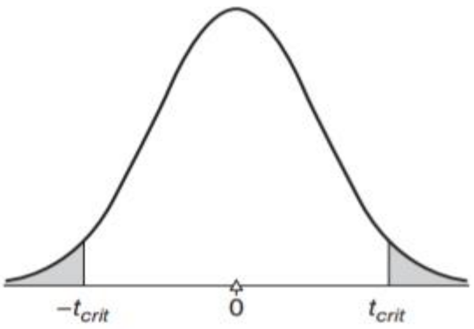
$$D = X_1 - X_2$$

  - To obtain the mean for a set of difference scores, add all difference scores and Divide by the number of scores, that is, where Dbar is the mean difference score, ΣD is the sum of all positive difference scores minus the sum of all negative difference scores, and n is the number of difference scores. The sign of Dbar is crucial.

- For example, in the current experiment, a positive value of Dbar would signify that EPO facilitates endurance, while a negative value of Dbar would signify that EPO hinders endurance.

- **COMPARING THE TWO EXPERIMENTS :**
  - To simplify comparisons, exactly the same six X1 scores and six X2 Scores in the Original EPO experiment with two independent samples are used to generate the six DScores in the new EPO experiment with repeated measures, as indicated in Table below.
  - Therefore, the sample mean difference also is the same, both for the original experiment, where 1 2 X X 11 6 5, and for the new experiment, where D 5.
  - To dramatize the beneficial effects of repeated measures, highly similar pairs of X1 and X2Scores appear in the new experiment.
  - For example, high endurance scores of 18 and 13 Minutes are paired, presumably for a very physically fit patient, while low scores of only5 and 3 minutes are paired, presumably for another patient in terrible shape.
  - Since in Real applications there is no guarantee that individual differences will be this large, The net effect of a repeated-measures experiment might not be as beneficial as that Described in the current analysis. Figure below shows the much smaller variability among paired differences in endurance scores, D, for the new experiment.
  - The range of scores in the top histogram for X1And X2 equals 15 (from 18 – 3), while that in the bottom histogram for D equals only 5 (from 7 – 2).
  - This suggests that once the new data have been analyzed with a t test for Repeated measures, it should be possible not only to reject the null hypothesis again, But also to claim a much smaller p-value than that ($p < .05$) for the t test for the original Experiment with two independent samples.

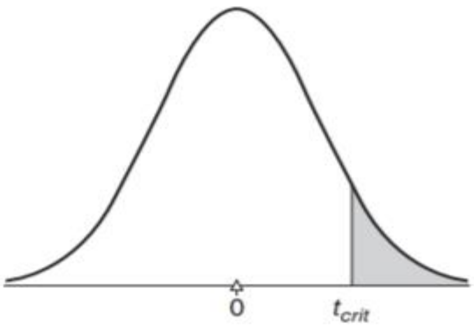
Table B[a]
CRITICAL VALUES OF *t*



Two-tailed or Nondirectional Test
LEVEL OF SIGNIFICANCE

$p > .05$ $p < .05$ $p < .01$ $p < .001$

df	.05*	.01**	.001
1	12.706	63.657	636.62
2	4.303	9.925	31.598
3	3.182	5.841	12.924
4	2.776	4.604	8.610
5	2.571	4.032	6.869
6	2.447	3.707	5.959
7	2.365	3.499	5.408
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
11	2.201	3.106	4.437
12	2.179	3.055	4.318
13	2.160	3.012	4.221
14	2.145	2.977	4.140
15	2.131	2.947	4.073
16	2.120	2.921	4.015
17	2.110	2.898	3.965
18	2.101	2.878	3.922
19	2.093	2.861	3.883
20	2.086	2.845	3.850
21	2.080	2.831	3.819
22	2.074	2.819	3.792
23	2.069	2.807	3.767
24	2.064	2.797	3.745
25	2.060	2.787	3.725
26	2.056	2.779	3.707
27	2.052	2.771	3.690
28	2.048	2.763	3.674
29	2.045	2.756	3.659
30	2.042	2.750	3.646
40	2.021	2.704	3.551
60	2.000	2.660	3.460
120	1.980	2.617	3.373
∞	1.960	2.576	3.291



One-tailed or Directional Test
LEVEL OF SIGNIFICANCE

$p > .05$ $p < .05$ $p < .01$ $p < .001$

df	.05	.01	.001
1	6.314	31.821	318.31
2	2.920	6.965	22.326
3	2.353	4.541	10.213
4	2.132	3.747	7.173
5	2.015	3.365	5.893
6	1.943	3.143	5.208
7	1.895	2.998	4.785
8	1.860	2.896	4.501
9	1.833	2.821	4.297
10	1.812	2.764	4.144
11	1.796	2.718	4.025
12	1.782	2.681	3.930
13	1.771	2.650	3.852
14	1.761	2.624	3.787
15	1.753	2.602	3.733
16	1.746	2.583	3.686
17	1.740	2.567	3.646
18	1.734	2.552	3.610
19	1.729	2.539	3.579
20	1.725	2.528	3.552
21	1.721	2.518	3.527
22	1.717	2.508	3.505
23	1.714	2.500	3.485
24	1.711	2.492	3.467
25	1.708	2.485	3.450
26	1.706	2.479	3.435
27	1.703	2.473	3.421
28	1.701	2.467	3.408
29	1.699	2.462	3.396
30	1.697	2.457	3.385
40	1.684	2.423	3.307
60	1.671	2.390	3.232
120	1.658	2.358	3.160
∞	1.645	2.326	3.090

[a] *Discussed in Section 13.2.*

**95% level of confidence.*

***99% level of confidence.*