



BITS Pilani
Pilani Campus

Self Study

Akanksha Bharadwaj
Asst. Professor, CS/IS Department

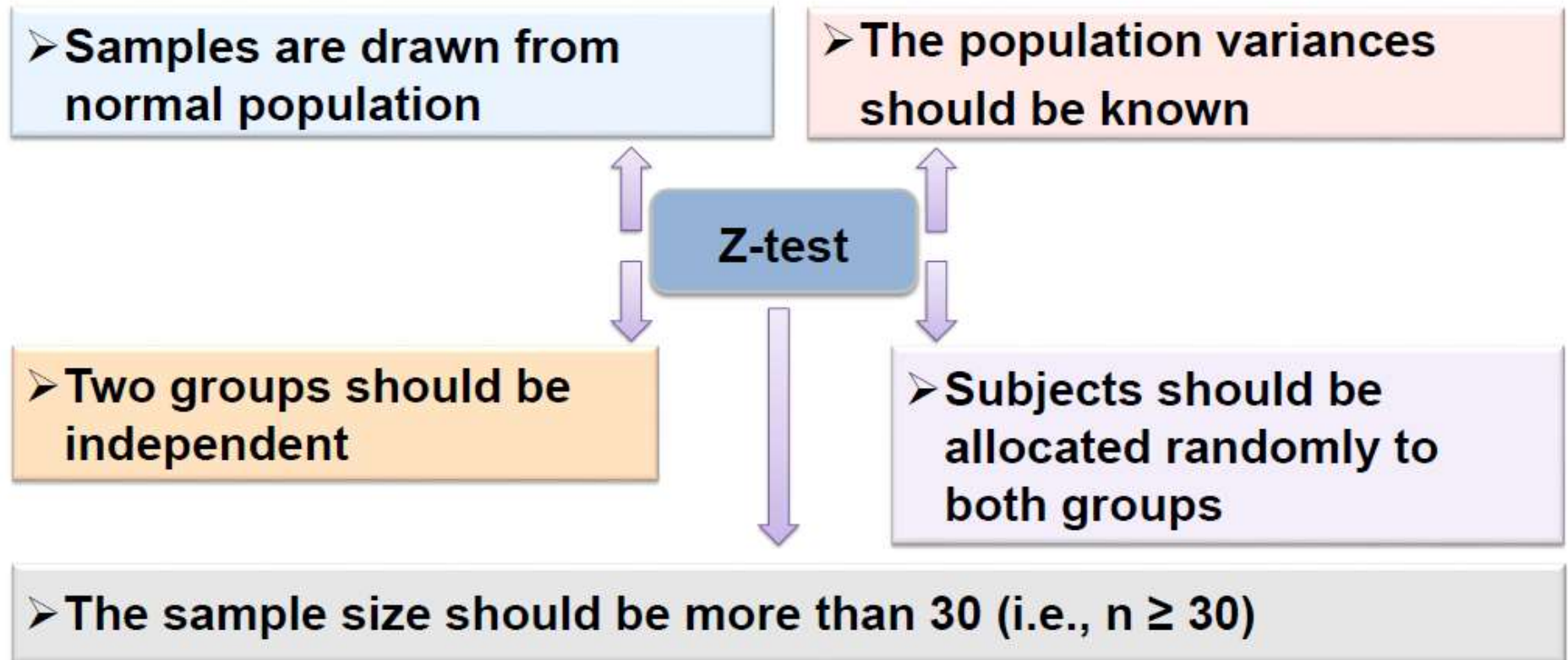


BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS Practice content

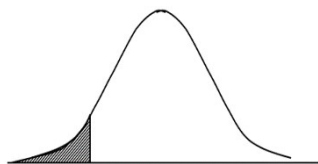
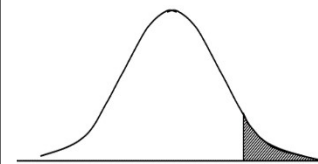
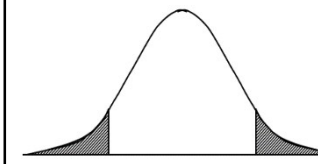
Assumptions for z test



Rejection criteria



Summary of One- and Two-Tail Tests

One-Tail Test (left tail)	One-Tail Test (right tail)	Two-Tail Test (Either left or right tail)
$H_0: \mu = \mu_0$ vs $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ vs $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$
		

Exercise



It is claimed that sports-car owners drive on the average 17000 kms per year. A consumer firm believes that the average mileage is probably higher. To check, the consumer firm obtained information from randomly selected 40 sports-car owners that resulted in a sample mean of 17352 kms with a population standard deviation of 1348 kms. At what can be concluded about this claim at

- (a) 5% level of significance (Critical value is 1.645)
- (b) 1% level of significance (Critical value is 2.331)

Solution



H_0



The average milage of sports-car as claimed and the sample average milage may be same

$$H_0 : \mu = \mu_0 = 17000$$

H_1



The average milage of sports-car as claimed may be **higher than** the sample average milage

$$H_1 : \mu > \mu_0 = 17000$$

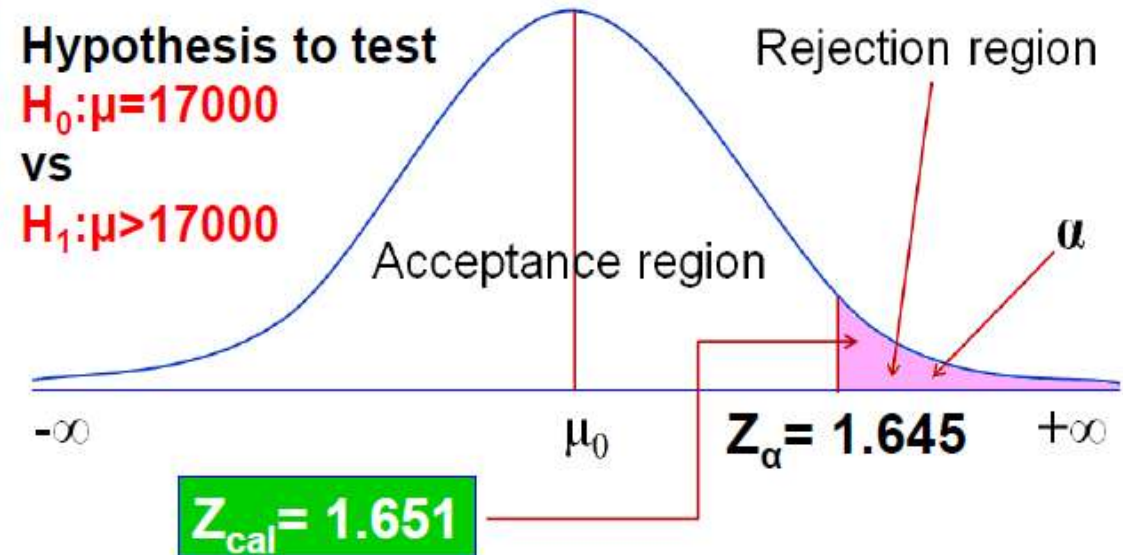
Solution



(a) At 5% level of significance with critical value 1.645

$$Z = \frac{17352 - 17000}{\frac{1348}{\sqrt{40}}} = 1.651$$

Critical value for
 $\alpha = 0.05$ is **1.645**
Since $Z = \mathbf{1.651} >$
1.645, Reject H_0
and Accept H_1



Solution



(b) At 1% level of significance with critical value 2.331

$$Z = \frac{17352 - 17000}{\frac{1348}{\sqrt{40}}} = 1.651$$

Critical value for

$\alpha = 0.01$ is **2.331**

Since $Z = 1.651$

< 2.330, Accept H_0

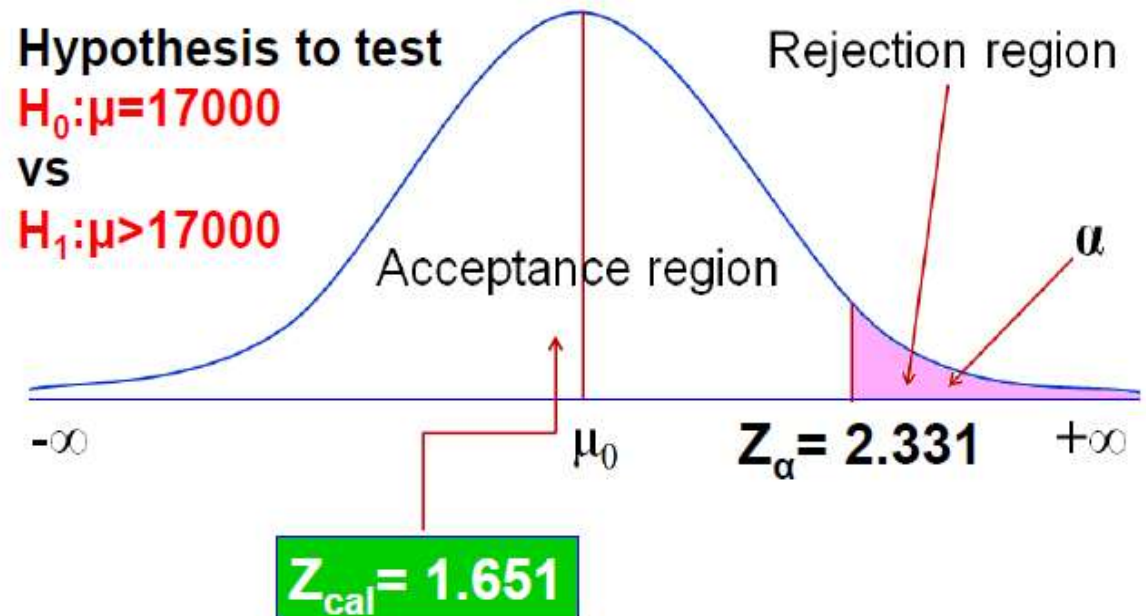
Reject H_1

Hypothesis to test

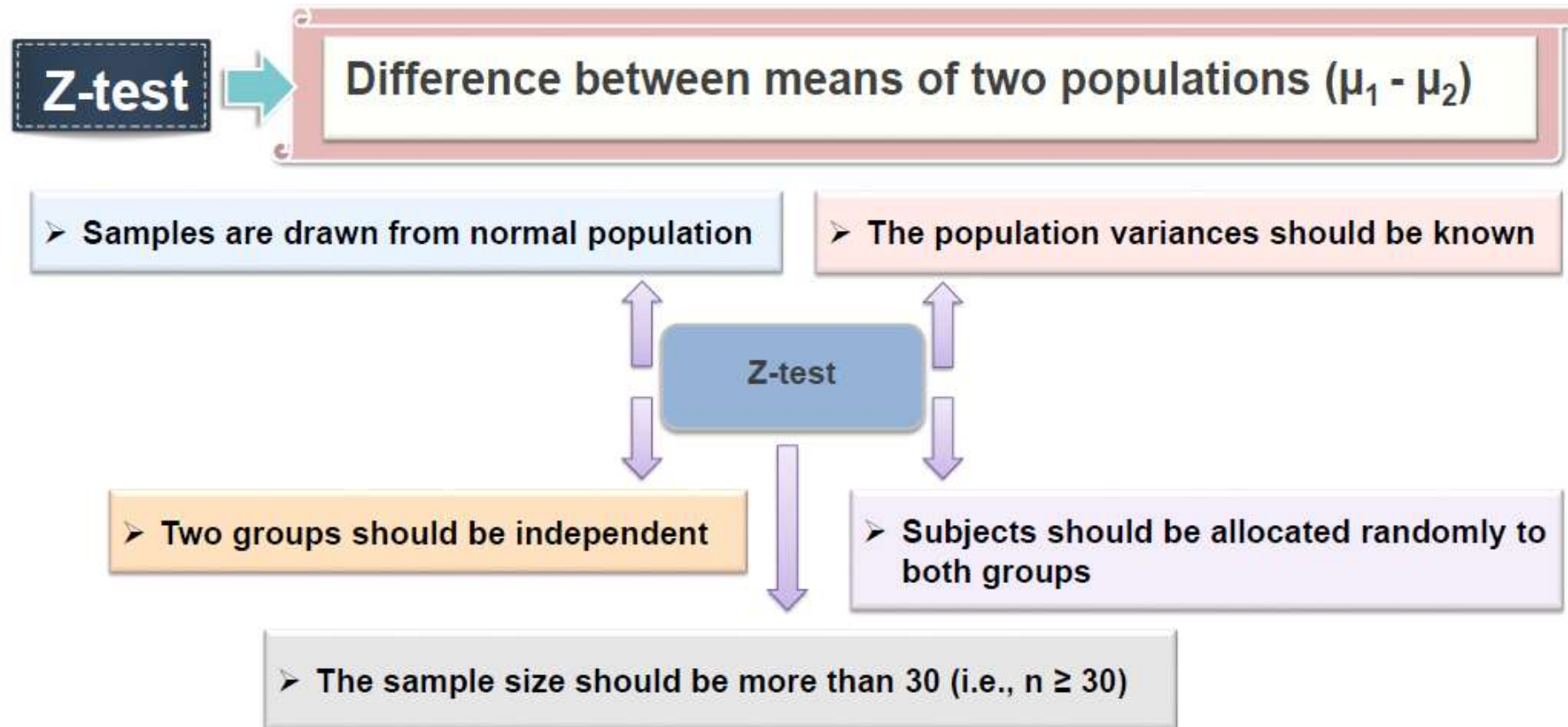
$H_0: \mu = 17000$

vs

$H_1: \mu > 17000$



Z test for difference between means



Exercise



The manager of a courier service believes that packets delivered at the beginning of the month are heavier than those delivered at the end of month. As an experiment, he weighed a random sample of 20 packets at the beginning of the month and found that the mean weight was 5.25 kg. A randomly selected 10 packets at the end of the month had a mean weight of 4.96 kg. It was observed from the past experience that the population variances are 1.20 kg and 1.15 kg. At 5% level of significance, can it be concluded that the packets delivered at the beginning of the month weigh more? Also find P-value and 95% confidence interval for the difference between the means.

Solution



At 5% (0.05) level of significance with critical value 1.645

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{5.25 - 4.96}{\sqrt{\frac{(1.20)^2}{20} + \frac{(1.15)^2}{10}}} = 0.642$$

Hypothesis to test

$$H_0: \mu_1 - \mu_2 = 0$$

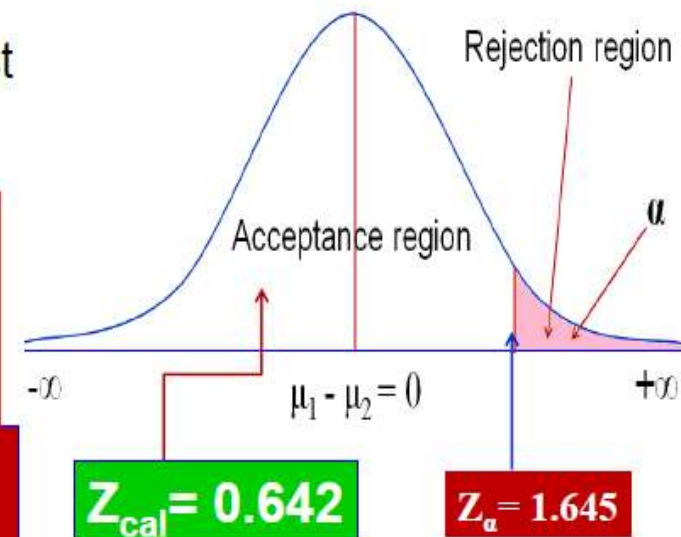
vs

$$H_1: \mu_1 - \mu_2 > 0$$

???

$$P(Z < 0.642) = 0.7389$$

95% CI for μ is
[- 4.54, 1.033]



Critical value for $\alpha = 0.05$ is 1.645. Since $Z = 0.642 < 1.645$, Accept H_0 Reject H_1

Confidence Intervals



- Sometimes being able to estimate the difference in the means of two populations is valuable.
- Algebraically, formula 10.1 can be manipulated to produce a formula for constructing confidence intervals for the difference in two population means.

CONFIDENCE INTERVAL TO
ESTIMATE $\mu_1 - \mu_2$ (10.2)

$$(\bar{x}_1 - \bar{x}_2) - z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example



Suppose a study is conducted to estimate the difference between middle-income shoppers and low-income shoppers in terms of the average amount saved on grocery bills per week by using coupons. Random samples of 60 middle-income shoppers and 80 low income shoppers are taken, and their purchases are monitored for one week. The average amounts saved with coupons, as well as sample sizes and population standard deviations are in the table below. This information can be used to construct a 98% confidence interval

Middle-Income Shoppers	Low-Income Shoppers
$n_1 = 60$	$n_2 = 80$
$\bar{x}_1 = \$5.84$	$\bar{x}_2 = \$2.67$
$\sigma_1 = \$1.41$	$\sigma_2 = \$0.54$

Solution



The z_c value associated with a 98% level of confidence is 2.33. This value, the data shown, and formula (10.2) can be used to determine the confidence interval.

$$\begin{aligned}(5.84 - 2.67) - 2.33\sqrt{\frac{1.41^2}{60} + \frac{0.54^2}{80}} &\leq \mu_1 - \mu_2 \leq (5.84 - 2.67) + 2.33\sqrt{\frac{1.41^2}{60} + \frac{0.54^2}{80}} \\ 3.17 - 0.45 &\leq \mu_1 - \mu_2 \leq 3.17 + 0.45 \\ 2.72 &\leq \mu_1 - \mu_2 \leq 3.62\end{aligned}$$

Exercise



- A consumer test group wants to determine the difference in gasoline mileage of cars using regular unleaded gas and cars using premium unleaded gas. Researchers for the group divided a fleet of 100 cars of the same make in half and tested each car on one tank of gas. Fifty of the cars were filled with regular unleaded gas and 50 were filled with premium unleaded gas. The sample average for the regular gasoline group was 21.45 miles per gallon (mpg), and the sample average for the premium gasoline group was 24.6 mpg. Assume that the population standard deviation of the regular unleaded gas population is 3.46 mpg, and that the population standard deviation of the premium unleaded gas population is 2.99 mpg. Construct a 95% confidence interval to estimate the difference in the mean gas mileage between the cars using regular gasoline and the cars using premium gasoline.

Solution

The z value for a 95% confidence interval is 1.96. The other sample information follows.

Regular	Premium
$n_r = 50$	$n_p = 50$
$\bar{x}_r = 21.45$	$\bar{x}_p = 24.6$
$\sigma_r = 3.46$	$\sigma_p = 2.99$

Based on this information, the confidence interval is

$$\begin{aligned}
 (21.45 - 24.6) - 1.96 \sqrt{\frac{3.46^2}{50} + \frac{2.99^2}{50}} &\leq \mu_1 - \mu_2 \leq (21.45 - 24.6) + 1.96 \sqrt{\frac{3.46^2}{50} + \frac{2.99^2}{50}} \\
 -3.15 - 1.27 &\leq \mu_1 - \mu_2 \leq -3.15 + 1.27 \\
 -4.42 &\leq \mu_1 - \mu_2 \leq -1.88
 \end{aligned}$$

We are 95% confident that the actual difference in mean gasoline mileage between the two types of gasoline is between -1.88 mpg and -4.42 mpg. The point estimate is -3.15 mpg.



The Difference In Two Means: Independent Samples And Population Variances Unknown

Introduction



- On many occasions, statisticians test hypotheses or construct confidence intervals about the difference in two population means and the population variances are not known. If the population variances are not known, the z methodology is not appropriate.

If $\sigma_1^2 = \sigma_2^2$, formula 10.1 algebraically reduces to

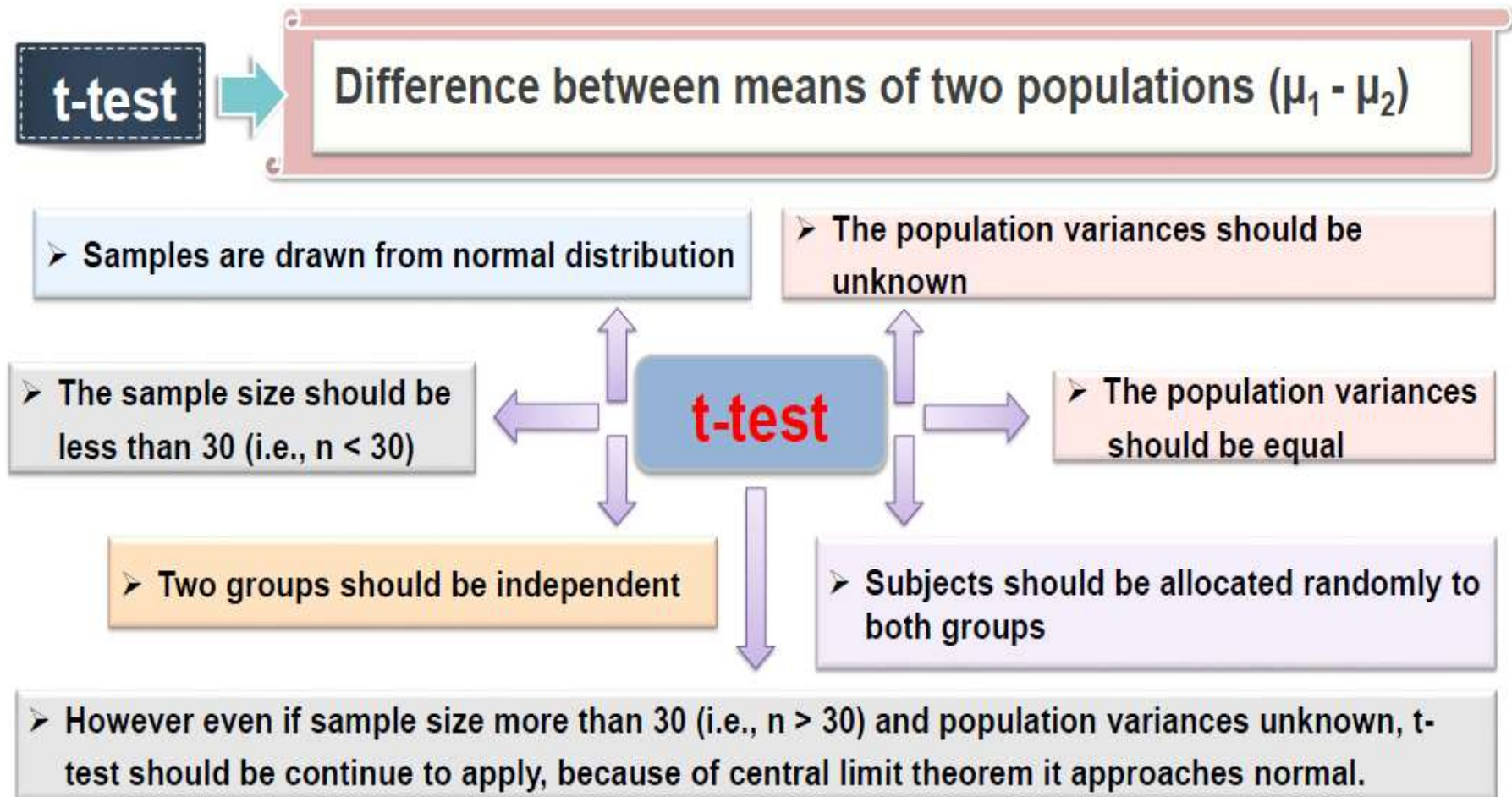
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

If σ is unknown, it can be estimated by *pooling* the two sample variances and computing a pooled sample standard deviation.

$$\sigma \approx s_p = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

s_p^2 is the weighted average of the two sample variances, s_1^2 and s_2^2 . Substituting this expression for σ and changing z to t produces a formula to test the difference in means.

Introduction



† FORMULA TO TEST THE
DIFFERENCE IN MEANS
ASSUMING σ_1^2, σ_2^2 ARE
EQUAL (10.3)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$df = n_1 + n_2 - 2$$

- When using formula 10.3 to test hypotheses about the difference in two means for small independent samples when the population variances are unknown, we must assume that the two samples come from populations in which the variances are essentially equal.

Note: If the equal variances assumption can not be met the following formula should be used.

† FORMULA TO TEST THE
DIFFERENCE IN MEANS

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1}}$$

Example



At the Hernandez Manufacturing Company, an application of this test arises. New employees are expected to attend a three-day seminar to learn about the company. At the end of the seminar, they are tested to measure their knowledge about the company. The traditional training method has been lecture and a question-and-answer session. Management decided to experiment with a different training procedure, which processes new employees in two days by using DVDs and having no question-and-answer session. If this procedure works, it could save the company thousands of dollars over a period of several years. However, there is some concern about the effectiveness of the two-day method, and company managers would like to know whether there is any difference in the effectiveness of the two training methods. To test the difference in the two methods, the managers randomly select one group of 15 newly hired employees to take the three-day seminar (method A) and a second group of 12 new employees for the two-day DVD method (method B).



Using $\alpha = .05$, the managers want to determine whether there is a significant difference in the mean scores of the two groups. They assume that the scores for this test are normally distributed and that the population variances are approximately equal.

Training Method A	Training Method B
56 50 52 44 52	59 54 55 65
47 47 53 45 48	52 57 64 53
42 51 42 43 44	53 56 53 57

Solution



HYPOTHESIZE:

STEP 1. The hypotheses for this test follow.

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_2: \mu_1 - \mu_2 \neq 0$$

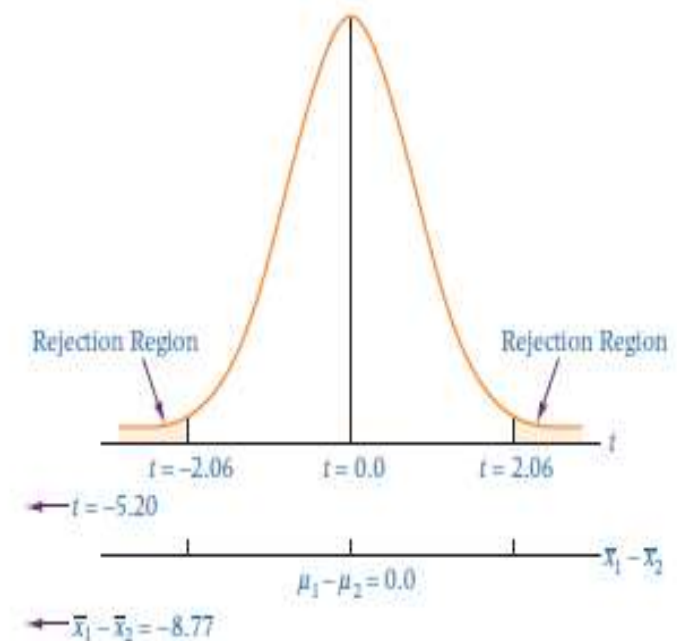
TEST:

STEP 2. The statistical test to be used is formula 10.3.

STEP 3. The value of alpha is .05.

STEP 4. Because the hypotheses are = and \neq , this test is two tailed. The degrees of freedom are 25 ($15 + 12 - 2 = 25$) and alpha is .05. The t table requires an alpha value for one tail only, and, because it is a two-tailed test, alpha is split from .05 to .025 to obtain the table t value: $t_{.025, 25} = \pm 2.060$.

The null hypothesis will be rejected if the observed t value is less than -2.060 or greater than $+2.060$.



STEP 5. The sample data are given in Table 10.2. From these data, we can calculate the sample statistics. The sample means and variances follow.

Method A	Method B
$\bar{x}_1 = 47.73$	$\bar{x}_2 = 56.5$
$s_1^2 = 19.495$	$s_2^2 = 18.273$
$n_1 = 15$	$n_2 = 12$

STEP 6. The observed value of t is

$$t = \frac{(47.73 - 56.50) - (0)}{\sqrt{\frac{(19.495)(14) + (18.273)(11)}{(15 + 12 - 2)}} \sqrt{\frac{1}{15} + \frac{1}{12}}} = -5.20$$

ACTION:

STEP 7. Because the observed value, $t = -5.20$, is less than the lower critical table value, $t = -2.06$, the observed value of t is in the rejection region. The null hypothesis is rejected. There is a significant difference in the mean scores of the two tests.

Exercise



Is there a difference in the way Chinese cultural values affect the purchasing strategies of industrial buyers in Taiwan and mainland China? A study by researchers at the National Chiao-Tung University in Taiwan attempted to determine whether there is a significant difference in the purchasing strategies of industrial buyers between Taiwan and mainland China based on the cultural dimension labeled “integration.” Integration is being in harmony with one’s self, family, and associates. For the study, 46 Taiwanese buyers and 26 mainland Chinese buyers were contacted and interviewed. Buyers were asked to respond to 35 items using a 9-point scale with possible answers ranging from no importance (1) to extreme importance (9). The resulting statistics for the two groups are shown in step 5. Using $\alpha = .01$, test to determine whether there is a significant difference between buyers in Taiwan and buyers in mainland China on integration. Assume that integration scores are normally distributed in the population.

Solution



HYPOTHESIZE:

STEP 1. If a two-tailed test is undertaken, the hypotheses and the table t value are as follows.

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_a: \mu_1 - \mu_2 \neq 0$$

TEST:

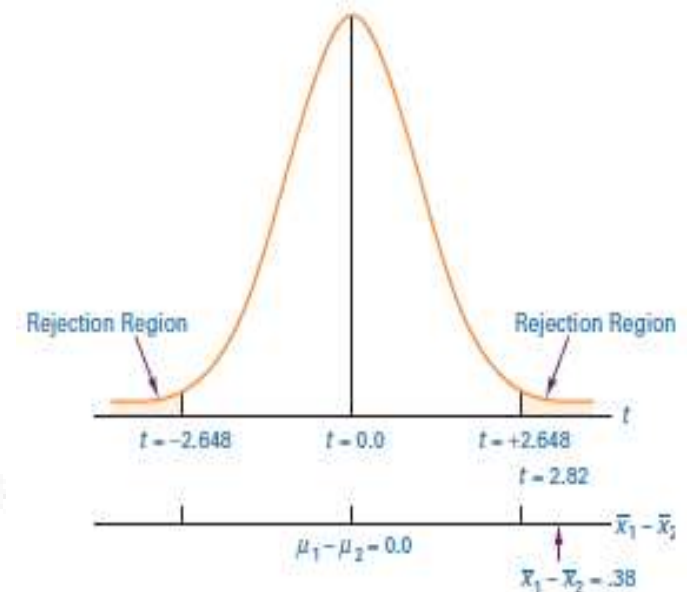
STEP 2. The appropriate statistical test is formula 10.3.

STEP 3. The value of alpha is .01.

STEP 4. The sample sizes are 46 and 26. Thus, there are 70 degrees of freedom.

With this figure and $\alpha/2 = .005$, critical table t values can be determined.

$$t_{.005, 70} = 2.648$$



STEP 5. The sample data follow.

Integration	
Taiwanese Buyers	Mainland Chinese Buyers
$n_1 = 46$	$n_2 = 26$
$\bar{x}_1 = 5.42$	$\bar{x}_2 = 5.04$
$s_1^2 = (.58)^2 = .3364$	$s_2^2 = (.49)^2 = .2401$
$df = n_1 + n_2 - 2 = 46 + 26 - 2 = 70$	

STEP 6. The observed t value is

$$t = \frac{(5.42 - 5.04) - (0)}{\sqrt{\frac{(.3364)(45) + (.2401)(25)}{46 + 26 - 2}} \sqrt{\frac{1}{46} + \frac{1}{26}}} = 2.82$$

ACTION:

STEP 7. Because the observed value of $t = 2.82$ is greater than the critical table value of $t = 2.648$, the decision is to reject the null hypothesis.

Confidence Intervals



- Confidence interval formulas can be derived to estimate the difference in the population means for independent samples when the population variances are unknown.

CONFIDENCE INTERVAL
TO ESTIMATE $\mu_1 - \mu_2$
ASSUMING THE
POPULATION VARIANCES
ARE UNKNOWN AND
EQUAL (10.4)

$$(\bar{x}_1 - \bar{x}_2) - t \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq$$
$$(\bar{x}_1 - \bar{x}_2) + t \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$
$$df = n_1 + n_2 - 2$$

Example



One group of researchers set out to determine whether there is a difference between “average Americans” and those who are “phone survey respondents.”* Their study was based on a well-known personality survey that attempted to assess the personality profile of both average Americans and phone survey respondents. Suppose they sampled nine phone survey respondents and 10 average Americans in this survey and obtained the results on one personality factor, conscientiousness, which are displayed in Table 10.3. Assume that conscientiousness scores are normally distributed in the population.

Phone Survey Respondents	Average Americans
35.38	35.03
37.06	33.90
37.74	34.56
36.97	36.24
37.84	34.59
37.50	34.95
40.75	33.30
35.31	34.73
35.30	34.79
	37.83
$n_1 = 9$	$n_2 = 10$
$\bar{x}_1 = 37.09$	$\bar{x}_2 = 34.99$
$s_1 = 1.727$	$s_2 = 1.253$
$df = 9 + 10 - 2 = 17$	

Solution



- The table t value for a 99% level of confidence and 17 degrees of freedom is $t_{.005,17} = 2.898$. The confidence interval is

$$(37.09 - 34.99) \pm 2.898 \sqrt{\frac{(1.727)^2(8) + (1.253)^2(9)}{9 + 10 - 2}} \sqrt{\frac{1}{9} + \frac{1}{10}}$$

$$2.10 \pm 1.99$$

$$0.11 \leq \mu_1 - \mu_2 \leq 4.09$$

Exercise



A coffee manufacturer is interested in estimating the difference in the average daily coffee consumption of regular-coffee drinkers and decaffeinated-coffee drinkers. Its researcher randomly selects 13 regular-coffee drinkers and asks how many cups of coffee per day they drink. He randomly locates 15 decaffeinated-coffee drinkers and asks how many cups of coffee per day they drink. The average for the regular-coffee drinkers is 4.35 cups, with a standard deviation of 1.20 cups. The average for the decaffeinated-coffee drinkers is 6.84 cups, with a standard deviation of 1.42 cups. The researcher assumes, for each population, that the daily consumption is normally distributed, and he constructs a 95% confidence interval to estimate the difference in the averages of the two populations.

Solution



The table t value for this problem is $t_{0.025, 26} = 2.056$. The confidence interval estimate is

$$\begin{aligned} & (4.35 - 6.84) \pm 2.056 \sqrt{\frac{(1.20)^2(12) + (1.42)^2(14)}{13 + 15 - 2}} \sqrt{\frac{1}{13} + \frac{1}{15}} \\ & -2.49 \pm 1.03 \\ & -3.52 \leq \mu_1 - \mu_2 \leq -1.46 \end{aligned}$$

The researcher is 95% confident that the difference in population average daily consumption of cups of coffee between regular- and decaffeinated-coffee drinkers is between 1.46 cups and 3.52 cups. The point estimate for the difference in population means is 2.49 cups, with an error of 1.03 cups.