Statistics & Probability

Chapter 10: STATISTICAL INFERENCE FOR TWO SAMPLES

FPT University

Department of Mathematics

Quy Nhon, 2023

Table of Contents

Table of Contents

Population 1 has mean μ_1 and variance σ_1^2, population 2 has mean μ_2 and variance σ_2^2. Inferences will be based on two random samples of sizes n_1 and n_2, respectively. That is,

- $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample of n_1 observations from population 1.
- $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample of n_2 observations from population 2.

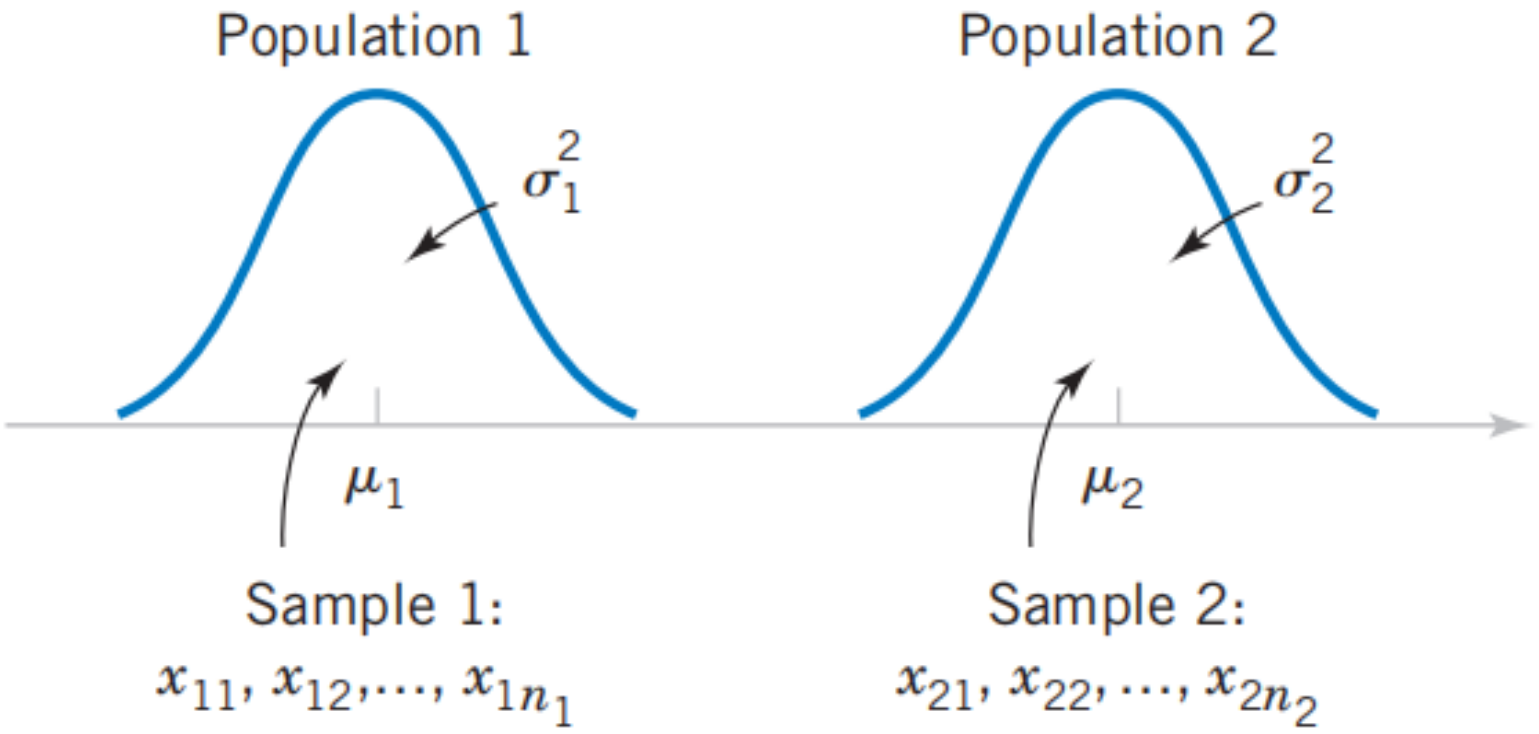


Figure. Two Independent Populations.

Assumptions for Two-Sample Inference

1. $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
2. $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2.
3. The two populations represented by X_1 and X_2 are independent.
4. Both populations are normal.

Remarks

- $\overline{X}_1 - \overline{X}_2$ is point estimator of $\mu_1 - \mu_2$.
- Based on the properties of expected values

$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2$$

and the variances of $\overline{X}_1 - \overline{X}_2$ is

$$V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

- The quantity

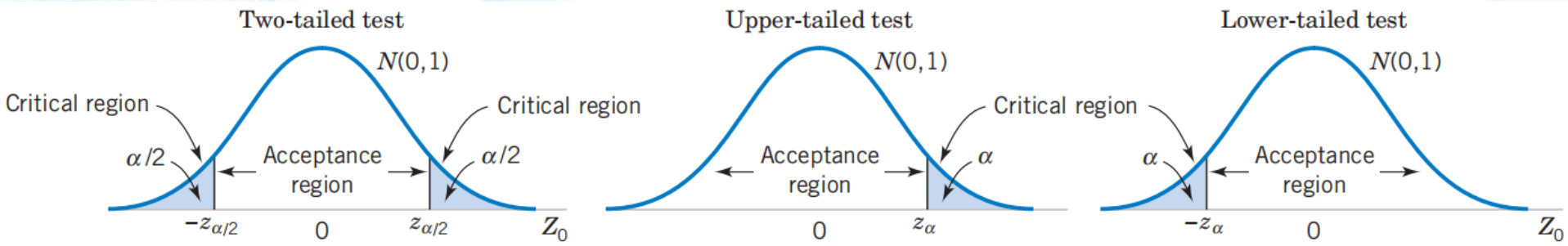
$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

Hypothesis Tests on the Difference in Means, Variances Known

Null hypothesis: H_0: $\mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (10\text{-}2)$$

Alternative Hypotheses	*P*-Value	Rejection Criterion For for Fixed-Level Tests
H_1: $\mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0\rvert$ and probability below $-\lvert z_0\rvert$, $P = 2[1 - \Phi(\lvert z_0\rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
H_1: $\mu_1 - \mu_2 > \Delta_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
H_1: $\mu_1 - \mu_2 < \Delta_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$


Two-tailed test
N(0,1)
Critical region
Critical region
α/2
Acceptance region
α/2
−z_{α/2}
0
z_{α/2}
Z_0
Upper-tailed test
N(0,1)
Critical region
Acceptance region
α
0
z_α
Lower-tailed test
N(0,1)
Critical region
α
Acceptance region
−z_α
0
Z_0

Example. A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample average drying times are $\overline{x}_1 = 121$ minutes and $\overline{x}_2 = 112$ minutes, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

Solution.

1. The null and alternative hypotheses are
$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 > 0.$$
2. Reject H_0 if the P-value is less than $\alpha = 0.05$.
3. Test statistic is
$$z_0 = \frac{\overline{x}_1 - \overline{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{121 - 112 - 0}{\sqrt{\frac{8^2}{10} + \frac{8^2}{10}}} = 2.52.$$

4. Since $z_0 = 2.52$, the P-value is

$$P\text{-value} = P(Z > 2.52) = 1 - \Phi(2.52) = 0.0059 < 0.05.$$

5. **Conclusion**: We reject H_0 at the $\alpha = 0.05$ level. We conclude that adding the new gradient to the paint significantly reduces the drying time.

Confidence Interval on the Difference in Means, variances Known

- $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample of n_1 observations from population 1.
- $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample of n_2 observations from population 2.
- $Z = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$ if the two populations are normal or is approximately standard normal if the conditions of the CLT apply. This implies

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha, \text{ or}$$

$$P\left(\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha.$$

Confidence Interval on the Difference in Means, Variances Known

Let $\overline{x}_1$ and $\overline{x}_2$ are the means of independent random samples of sizes n_1 and n_2 from two independent normal populations with known variances σ_1^2 and σ_2^2, respectively. Then, a $100(1-\alpha)\%$ **confidence interval** for $\mu_1 - \mu_2$ is

$$\overline{x}_1 - \overline{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \overline{x}_1 - \overline{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

Example (Aluminum Tensile Strength)

Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows:

$$n_1 = 10,\ \overline{x}_1 = 87.6,\ \sigma_1 = 1 \quad \text{and} \quad n_2 = 12,\ \overline{x}_2 = 74.5,\ \sigma_2 = 1.5.$$

Construct 90% confidence interval on the difference in means.

Solution. Let μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars. Then,

$$\overline{x}_1 - \overline{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \overline{x}_1 - \overline{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$87.6 - 74.5 - 1.645\sqrt{\frac{1^2}{10} + \frac{1.5^2}{12}} \le \mu_1 - \mu_2 \le 87.6 - 74.5 + 1.645\sqrt{\frac{1^2}{10} + \frac{1.5^2}{12}}$$

$$12.22 \le \mu_1 - \mu_2 \le 13.98.$$

Note. The confidence interval does not include zero, implying that the mean strength of aluminum grade 1 (μ_1) exceeds the mean strength of aluminum grade 2 (μ_2). In fact, we are 90% confident that the mean tensile strength of aluminum grade 1 exceeds that of aluminum grade 2 by between 12.22 and 13.98 kilograms per square millimeter.

Choice of Sample Size

Assume

- standard deviations σ_1 and σ_2 are known (at least approximately),
- two sample sizes n_1 and n_2 are equal, $n_1 = n_2 = n$.

Then, we can determine the same size required that the error in estimating $\mu_1 - \mu_2$ by $\overline{x}_1 - \overline{x}_2$ will be less than E at $100(1-\alpha)\%$ confidence.

Sample Size for a Confidence Interval on the Difference in Means, σ_1, σ_2 Known

$$n = \left\lceil \left(\frac{z_{\alpha/2}}{E} \right)^2 (\sigma_1^2 + \sigma_2^2) \right\rceil .$$

One-Sided Confidence Bound on the Difference in Means

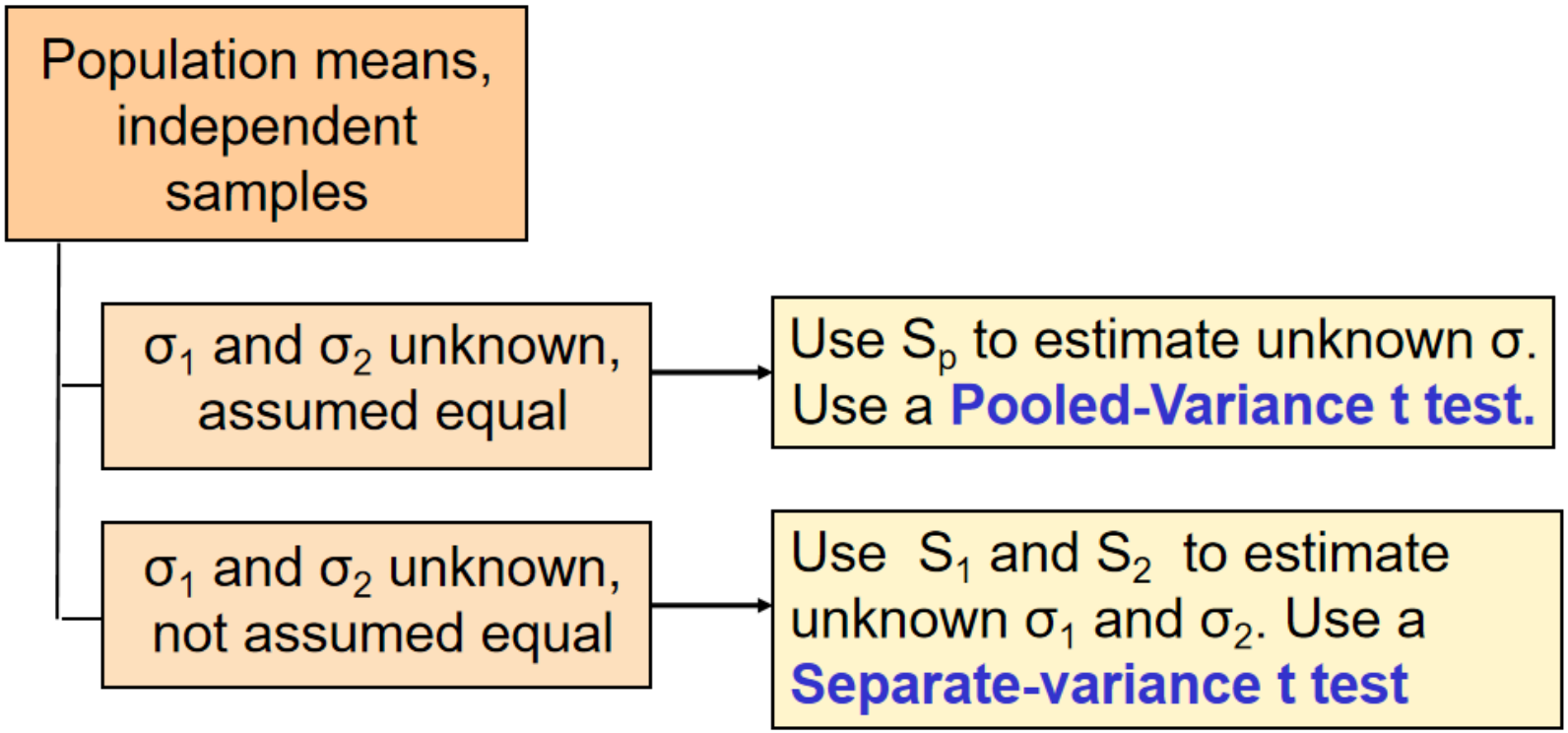
1. A $100(1-\alpha)\%$ **upper**-confidence bound for $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \leq \overline{x}_1 - \overline{x}_2 + z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

2. A $100(1-\alpha)\%$ **lower**-confidence bound for $\mu_1 - \mu_2$ is

$$\overline{x}_1 - \overline{x}_2 - z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2.$$

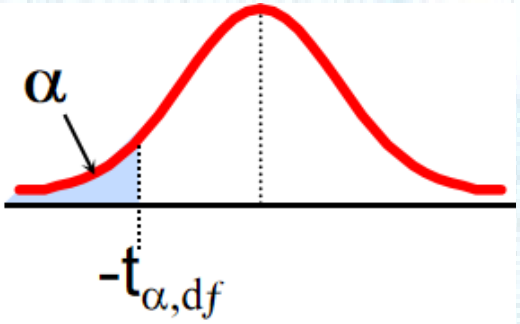
Table of Contents


Population means, independent samples
σ_1 and σ_2 unknown, assumed equal
Use S_p to estimate unknown σ. Use a **Pooled-Variance t test.**
σ_1 and σ_2 unknown, not assumed equal
Use S_1 and S_2 to estimate unknown σ_1 and σ_2. Use a **Separate-variance t test**

Two Population Means, Independent Samples

Lower-tail test

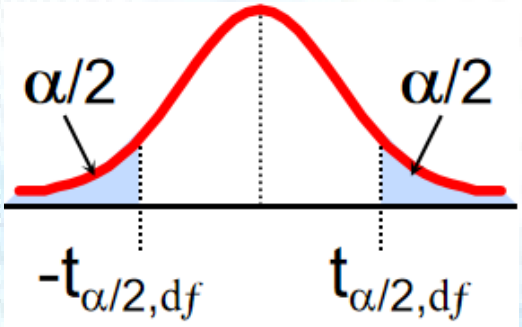
$$H_0 : \mu_1 - \mu_2 = \Delta_0$$
$$H_1 : \mu_1 - \mu_2 < \Delta_0$$



Reject H_0 if $t_0 < -t_{\alpha,df}$

Two-tail test

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$
$$H_1 : \mu_1 - \mu_2 \neq \Delta_0$$



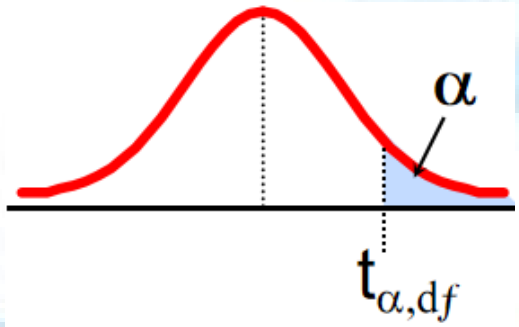
Reject H_0 if

$t_0 < -t_{\alpha/2,df}$

or $t_0 > t_{\alpha/2,df}$

Upper-tail test

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$
$$H_1 : \mu_1 - \mu_2 > \Delta_0$$



Reject H_0 if $t_0 > t_{\alpha,df}$

Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$, Pooled-Variance t Test

Assumptions:

- Samples are randomly and independently drawn.
- Populations are normally distributed or both sample sizes are at least 30.
- Population variances are unknown but assumed equal.

Then, we can use a **pooled-variance t test**.

$$T_0 = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

- $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$: pooled estimator of σ^2 (or pooled variance)
- $\overline{X}_1, \overline{X}_2$: means of the samples taken from populations 1 and 2, respectively
- S_1^2, S_2^2: variances of the samples taken from populations 1 and 2, respectively
- n_1, n_2: sizes of the samples taken from populations 1 and 2, respectively.

Note. The T_0 test statistic follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Remarks

The pooled estimator S_p^2 can be written as

$$S_p^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2 = wS_1^2 + (1 - w)S_2^2,$$

where $0 \leq w \leq 1$, is a **weighted average** of the two sample variances S_1^2 and S_2^2.

Test of Hypotheses for Difference in Means (Pooled-Variance t Test)

S_1. Construct the two hypotheses $H_0 : \mu_1 - \mu_2 = \Delta_0$ and $H_1 : \mu_1 - \mu_2 \neq \Delta_0$.

S_2. Identify acceptance region, use t-distribution with $df = n_1 + n_2 - 2$.

S_3. Determine the test statistic $t_0 = \dfrac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{s_p^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)}}$.

S_4. Make a decision:

+ If the test statistic t_0 is in critical region, then reject H_0.
+ If the test statistic t_0 is in acceptance region, then fail to reject H_0.

Example. Assume the samples are from normal populations having equal variances which have the data related to sales location as follows

- **Special Front**: $224,\ 189,\ 248,\ 285,\ 273,\ 190,\ 243,\ 215,\ 280,\ 317$
- **In-Aisle**: $192,\ 236,\ 164,\ 154,\ 189,\ 220,\ 261,\ 186,\ 219,\ 202.$

Determine whether there is a difference between the means. Use $\alpha = 0.05$.

1. The null and alternative hypotheses are
$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0.$$
2. We have $\alpha = 0.05$ and $df = n_1 + n_2 - 2 = 18$, so acceptance region is
$$-t_{0.025,18} = -2.1009 \leq t_0 \leq +2.1009 = t_{0.025,18}.$$
3. $\overline{x}_1 = 246.4,\ n_1 = 10,\ \overline{x}_2 = 202.3,\ n_2 = 10$. Here,
$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{9 \cdot 42.5420^2 + 9 \cdot 32.5271^2}{9 + 9} = 1433.9167, \text{ and}$$
$$t_0 = \frac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(246.4 - 202.3) - 0.0}{\sqrt{1433.9167\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{44.1}{\sqrt{286.7833}} = 2.6041.$$
4. Since $t_0 = 2.6041 > +2.1009 = t_{0.025,18}$, we reject H_0.

FUQN

t-Test: Two-Sample Assuming Equal Variances		
	Special Front	In-Aisle
Mean	246.4	202.3
Variance	1809.822222	1058.011111
Observations	10	10
Pooled Variance	1433.916667	
Hypothesized Mean Difference	0	
df	18	
t Stat	2.604123851	
P(T<=t) one-tail	0.008971549	
t Critical one-tail	1.734063607	
P(T<=t) two-tail	0.017943098	
t Critical two-tail	2.10092204	

Result	**Conclusions**
The $t_0 = 2.6041 > 2.1009 = t_{0.025,9}$.	1. Reject the null hypothesis H_0.
	2. Conclude that evidence exists that the mean sales are different for the two sales locations.
The t test P-value $= 0.0179 < \alpha = 0.05$.	3. The probability of observing a difference in the two sample means this large or larger is 0.0179.
The t_0 is positive.	4. Conclude that the mean sales are higher for the special front location.

Recall

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

Confidence Interval on the Difference in Means, Variances Unknown and Equal

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{x}_1 - \overline{x}_2) - t_{\alpha/2,\mathtt{df}}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \le \mu_1 - \mu_2 \le (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2,\mathtt{df}}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}.$$

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (10\text{-}14)$$

Alternative Hypothesis	*P-Value*	Rejection Criterion for Fixed-Level Tests
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert t_0\rvert$ and probability below $-\lvert t_0\rvert$	$t_0 > t_{\alpha/2,n_1+n_2-2}$ or $t_0 < -t_{\alpha/2,n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probability above t_0	$t_0 > t_{\alpha,n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probability below t_0	$t_0 < -t_{\alpha,n_1+n_2-2}$

Quiz

Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, it should be adopted, providing it does not change the process yield. A test is run in the pilot plant and results in the data shown in Table. Is there any difference between the mean yields? Use $\alpha = 0.05$, and assume equal variances.

Observation number	**Catalyst 1**	**Catalyst 2**
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75
	$\overline{x}_1 = 92.255$	$\overline{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

Case 2: $\sigma_1^2 \neq \sigma_2^2$, Separate-Variance t Test

Test Statistic for the Difference in Means, Variances Unknown, Not Assumed Equal

If $H_0 : \mu_1 = \mu_2$ is true, the statistic

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

is distributed approximately as t with degrees of freedom given by

$$\nu = \left\lfloor \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \right\rfloor .$$

If $\sigma_1^2 \neq \sigma_2^2$, the hypotheses on differences in the means of two normal distributions are tested as in the equal variances case, except that

- T_0^* is used as the test statistic, and
- $df = \nu$.

Approximate Confidence Interval on the Difference in Means, Variances Unknown and Not Assumed Equal

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{x}_1 - \overline{x}_2) - t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Example

Read more in Textbook page 257.

Table of Contents

Large-Sample Tests on the Difference in Population Proportions

Assume that two independent random samples of sizes n_1 and n_2 (large enough) are taken from two populations. Let X_1 and X_2 represent the number of observations that belong to the samples 1 and 2, respectively. The estimators of the population proportions $P_1 = \dfrac{X_1}{n_1}$ and $P_2 = \dfrac{X_2}{n_2}$ have approximately normal distributions.

1. Sample proportions: $\hat{p}_1 = \dfrac{x_1}{n_1}$ and $\hat{p}_2 = \dfrac{x_2}{n_2}$.
2. $\hat{p}_1 - \hat{p}_2$ is point estimator of $p_1 - p_2$.
3. If n_1 and n_2 are large enough, we have

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2; \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

4. Test statistic for the difference of two population proportions

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \sim N(0,1).$$

5. Pooled proportion $\overline{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$.

Population Proportions

Lower-tail test

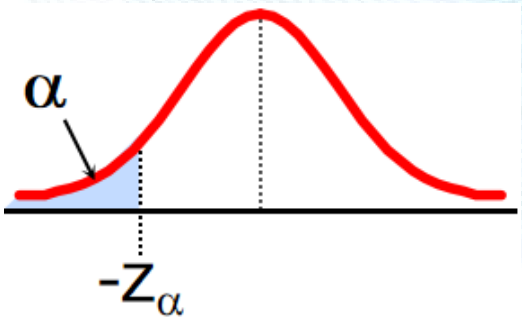
$H_0 : p_1 = p_2$

$H_1 : p_1 < p_2$

or

$H_0 : p_1 - p_2 = 0$

$H_1 : p_1 - p_2 < 0$



Reject H_0 **if** $z_0 < -z_\alpha$

Two-tail test

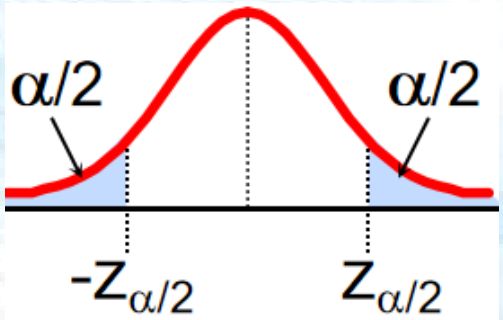
$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

or

$H_0 : p_1 - p_2 = 0$

$H_1 : p_1 - p_2 \neq 0$



Reject H_0 **if**

$z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$

Upper-tail test

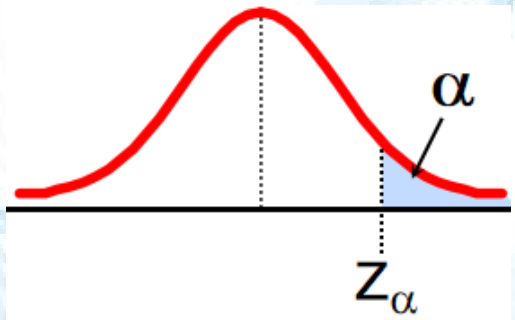
$H_0 : p_1 = p_2$

$H_1 : p_1 > p_2$

or

$H_0 : p_1 - p_2 = 0$

$H_1 : p_1 - p_2 > 0$



Reject H_0 **if**

$z_0 > z_\alpha$

FUQN

Pooled proportion $\overline{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$.

Test of Hypotheses for Difference in Proportions

S_1. Construct the two hypotheses $H_0 : p_1 - p_2 = 0$ and $H_1 : p_1 - p_2 \neq 0$.

S_2. Determine the test statistic $z_0 = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$.

S_3. Identify acceptance region, use $Z \sim N(0,1)$.

S_4. Make a decision:

+ If the test statistic is in critical region, then reject H_0.
+ If the test statistic is in acceptance region, then fail to reject H_0.

Is there a significant difference between the proportion of men and the proportion of women who will vote Yes on Proposition A? In two random samples, 36 of 72 men and 35 of 50 women indicated they would vote Yes. Test at the 0.05 level of significance.

- The hypothesis test is

$$H_0 : p_1 - p_2 = 0 \text{ (the two proportions are equal)}$$

$$H_1 : p_1 - p_2 \neq 0 \text{ (there is a significant difference between proportions).}$$

- The sample proportions are

$$\text{Men : } \hat{p}_1 = \frac{36}{72} = 0.5 \quad \text{and} \quad \text{Women : } \hat{p}_2 = \frac{35}{50} = 0.7.$$

- The pooled estimate for the overall proportion is

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{36 + 35}{72 + 50} = 0.582.$$

- Test statistic is

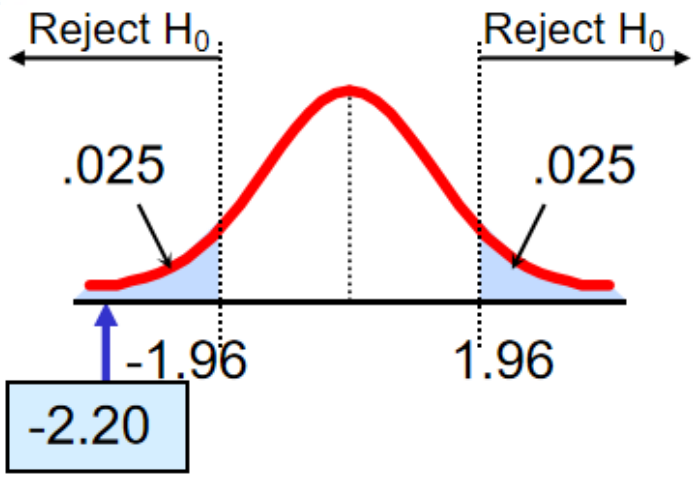
$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.5 - 0.7) - 0}{\sqrt{0.582(1-0.582)\left(\frac{1}{72} + \frac{1}{50}\right)}} = -2.20.$$

- Since $\alpha = 0.05$, the critical values are $\pm z_{\alpha/2} = \pm 1.96$ and nonrejection region is

$$-1.96 \leq z_0 \leq +1.96.$$

FUQN

- Since $-2.20 < -1.96$, we reject H_0.



- **Conclusion**: There is evidence of a significant difference in the proportion of men and women who will vote yes.

Approximate Confidence Interval on the Difference in Population Proportions

A $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

Is there a significant difference between the proportion of men and the proportion of women who will vote Yes on Proposition A? In two random samples, 36 of 72 men and 35 of 50 women indicated they would vote Yes. Construct 95% confidence interval for the difference between the proportions.

- We have
$$x_1 = 36,\ n_1 = 72,\ x_2 = 35,\ n_2 = 50.$$
- Since $1 - \alpha = 0.95$, then $z_{\alpha/2} = 1.96$,
- The sample proportions are
$$\text{Men}: \ \hat{p}_1 = \frac{36}{72} = 0.5 \quad \text{and} \quad \text{Women}: \ \hat{p}_2 = \frac{35}{50} = 0.7.$$
- The 95% confidence interval for the difference between the proportions is
$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$
$$= (0.5 - 0.7) \pm 1.96\sqrt{\frac{0.5 \cdot 0.5}{72} + \frac{0.7 \cdot 0.3}{50}} = (-0.37, -0.03).$$
- This interval does not contains 0, so we can be 95% confident the two proportion are different.

Recall

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2}, \quad \hat{p}_1 = \frac{x_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{x_2}{n_2}.$$

And

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Quiz

Let $n_1 = 100,\ X_1 = 45,\ n_2 = 50$ and $X_2 = 25$.

1. At the 0.01 level of significance, is there evidence of a significant difference between the two population proportions?
2. Construct a 99% confidence interval estimate for the difference between the two population proportions.

Quiz (St. John's Wort)

Extracts of St. John's Wort are widely used to treat depression. An article in the April $18,\ 2001$, issue of the Journal of the American Medical Association ("Effectiveness of St. John's Wort on Major Depression: A Randomized Controlled Trial") compared the efficacy of a standard extract of St. John's Wort with a placebo in 200 outpatients diagnosed with major depression. Patients were randomly assigned to two groups; one group received the St. John's Wort, and the other received the placebo. After eight weeks, 19 of the placebo-treated patients showed improvement, and 27 of those treated with St. John's Wort improved.

1. Is there any reason to believe that St. John's Wort is effective in treating major depression? Use $\alpha = 0.05$.
2. Construct 95% confidence interval for difference of two these proportions.

Thank you!