Chapter 10:
Statistical Inference for Two Samples

MAS291 - STATISTICS & PROBABILITY

Ly Anh Duong

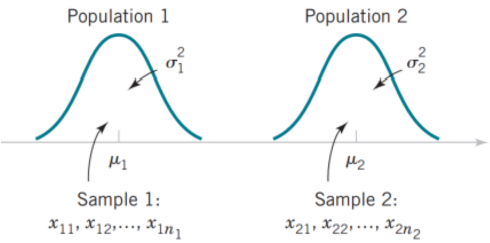
duongla3@fe.edu.vn

Table of Contents

Inference on the Difference in Means of Two Normal Distributions, Variances Known

Assumptions for Two-Sample Inference

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



Two independent populations

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2,\ V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Tests on the Difference in Means, σ Known

Null hypothesis:	$H_0 : \mu_1 - \mu_2 = \Delta_0$	
Test statistic:	$Z_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	
Alternative Hypotheses	**P-Value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0 \rvert$ and 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, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

Exercise

1 Inference on the Difference in Means of Two Normal Dist., Variances Known

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:

$$\overline{x}_1 = 121\,\text{minutes} \quad \text{and} \quad \overline{x}_2 = 112\,\text{minutes}.$$

What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

Solution

1. Establishing

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

2. Test statistic

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{121 - 112}{\sqrt{\frac{8^2}{10} + \frac{8^2}{10}}} = 2.52$$

3. Conclusion
   $z_0 = 2.52 > z_{0.05} \Rightarrow$ reject H_0

Confidence Interval on a Difference in Means, Variances known

If $\overline{x}_1$ 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, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

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

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

Exercise

1 Inference on the Difference in Means of Two Normal Dist., Variances Known

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:

$$n_1 = 10,\ \bar{x}_1 = 87.6,\ \sigma_1 = 1,\ n_2 = 12,\ \bar{x}_2 = 74.5,\ \sigma_2 = 1.5$$

If μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars, we may find a 90% CI on the difference in mean strength $\mu_1 - \mu_2$ as follows:

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

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

Therefore, the 90% confidence interval on the difference in mean tensile strength is

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

Choice of Sample Size & One-Sided Confidence Bounds

- **Choice of Sample Size:**

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

- **One-Sided Upper-Confidence Bound:**

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

- **One-Sided Lower-Confidence Bound:**

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

Table of Contents

Case 1: Variance unknown and $\sigma_1 = \sigma_2 = \sigma$

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

Test statistic: $T_0 = \dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

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

t distribution with $n_1 + n_2 - 2$ degrees of freedom

Exercise

Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently used; but catalyst 2 is acceptable. Because catalyst 2 is cheaper, it should be adopted, if it does not change the process yield. A test is run in the pilot plant and results in the data shown in Table 10.1. Figure 10.2 presents a normal probability plot and a comparative box plot of the data from the two samples.

Is there any difference in 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
	$\bar{x}_1 = 92.255$	$\bar{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

1. Establishing

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

2. Test statistic

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{2}{n_2}}}, s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{7(2.39)^2 + 7(2.98)^2}{8 + 8 - 2} = 7.30$$

So $s_p = \sqrt{7.30} = 2.70$ and $t_0 = \dfrac{92.2555 - 92.733}{2.7\sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.35$

3. Conclusion
$|t_0| = 0.35$

Case 2: Variances Unknown, small sample and $\sigma_1 \neq \sigma_2$

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

Test statistic: $T_0^* = \dfrac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \quad \nu = \dfrac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$

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 below $-\lvert t_0 \rvert$	$t_0 > t_{\alpha/2,\nu}$ or $t_0 < -t_{\alpha/2,\nu}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	Probability above t_0	$t_0 > t_{\alpha,\nu}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	Probability below t_0	$t_0 < -t_{\alpha,\nu}$

t distribution with ν degrees of freedom.

If ν is not an integer, round down to the nearest integer.

Exercise

Arsenic concentration in public drinking water supplies is a potential health risk. An article in the Arizona Republic (May 27, 2001) reported drinking water arsenic concentrations in parts per billion (ppb) for 10 metropolitan Phoenix communities and 10 communities in rural Arizona. We wish to determine whether any difference exists in mean arsenic concentrations for metropolitan Phoenix communities and for communities in rural Arizona

The data follow: ...

Metro Phoenix ($\bar{x}_1 = 12.5$, $s_1 = 7.63$)	Rural Arizona ($\bar{x}_2 = 27.5$, $s_2 = 15.3$)
Phoenix, 3	Rimrock, 48
Chandler, 7	Goodyear, 44
Gilbert, 25	New River, 40
Glendale, 10	Apache Junction, 38
Mesa, 15	Buckeye, 33
Paradise Valley, 6	Nogales, 21
Peoria, 12	Black Canyon City, 20
Scottsdale, 25	Sedona, 12
Tempe, 15	Payson, 1
Sun City, 7	Casa Grande, 18

Solution

1. Establishing

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

2. Test statistic

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{12.5 - 27.5}{\sqrt{\frac{(7.63)^2}{10} + \frac{(15.3)^2}{10}}} = -2.77$$

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left[\frac{(7.63)^2}{10} + \frac{(15.3)^2}{10}\right]^2}{\frac{\left[\frac{(7.63)^2}{10}\right]^2}{9} + \frac{\left[\frac{(15.3)^2}{10}\right]^2}{9}} = 13.2 \approx 13$$

3. Conclusion $t_0^* = -2.77 < -t_{0.025,13} = -2.160 \Rightarrow$ reject H_0

Case 1: CI on the difference in means, variance unknown and equal

If $\bar{x}_1$, $\bar{x}_2$, s_1^2, and s_2^2 are the sample means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown but equal variances, a $100(1-\alpha)\%$ confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2, n_1+n_2-2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, n_1+n_2-2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ is the pooled estimate of the common population standard deviation, and $t_{\alpha/2, n_1+n_2-2}$ is the upper $\alpha/2$ percentage point of the t-distribution with n_1+n_2-2 degrees of freedom.

Exercise

An article in the journal Hazardous Waste and Hazardous Materials (1989, Vol. 6) reported the results of an analysis of the weight of calcium in standard cement and cement doped with lead. Reduced levels of calcium would indicate that the hydration mechanism in the cement is blocked and would allow water to attack various locations in the cement structure. Ten samples of standard cement had an average weight percent calcium of $\bar{x}_1 = 90.0$ with a sample standard deviation of $s_1 = 5.0$, and 15 samples of the lead-doped cement had an average weight percent calcium of $\bar{x}_2 = 87.0$ with a sample standard deviation of $s_2 = 4.0$.

Solution

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{9(5.0)^2 + 14(4.0)^2}{10+15-2} = 19.52$$

$$\bar{x}_1 - \bar{x}_2 - t_{0.025,23} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{0.025,23} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$90.0 - 87.0 - 2.069(4.4)\sqrt{\frac{1}{10} + \frac{1}{15}} \le \mu_1 - \mu_2 \le 90.0 - 87.0 + 2.069(4.4)\sqrt{\frac{1}{10} + \frac{1}{15}}$$

which reduces to

$$-0.72 \le \mu_1 - \mu_2 \le 6.72$$

Case 2: Approximate CI on the Difference in Means, Variances Unknown and Not Assumed Equal

If $\bar{x}_1, \bar{x}_2, s_1^2$, and s_2^2 are the means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown and unequal variances, an approximate $100(1-\alpha)\%$ confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $\nu = \dfrac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$ and $t_{\alpha/2,v}$ is the upper $\alpha/2$ percentage point of the t-distribution with v degrees of freedom.

Table of Contents

Large-Sample Test on the Difference in Population Proportions

3 Inference on Two Population Proportions

Null hypothesis: $H_0: p_1 = p_2$

Test statistic: $$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Alternative Hypothesis	**P-Value**	**Rejection Criterion for Fixed-Level Tests**
$H_1: p_1 \neq p_2$	$P = 2[1 - \Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: p_1 > p_2$	$P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: p_1 < p_2$	$P = \Phi(z_0)$	$z_0 < -z_\alpha$

Exercise

3 Inference on Two Population Proportions

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 8 weeks, 19 of the placebo-treated patients showed improvement, and 27 of those treated with St. John's Wort improved. Is there any reason to believe that St. John's Wort is effective in treating major depression? Use $\alpha = 0.05$

Solution

3 Inference on Two Population Proportions

1. Establishing: $H_0 : p_1 = p_2, \ H_1 : p_1 > p_2$
2. Test statistic

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

   where $\hat{p}_1 = \frac{27}{100} = 0.27, \hat{p}_2 = \frac{19}{100} = 0.19, n_1 = n_2 = 100,$ and
   $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \frac{19+27}{100+100} = 0.23$
   Compute

$$z_0 = \frac{0.27 - 0.19}{\sqrt{0.23(0.77)(\frac{1}{100} + \frac{1}{100})}} = 1.34$$

3. Conclusion: cannot reject H_0 since $z_0 = 1.34 < z_{0.05}$

Confidence Interval on the Difference in Population Proportions

If $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of observations in two independent random samples of sizes n_1 and n_2 that belong to a class of interest, an approximate two-sided $100(1-\alpha)\%$ confidence interval on the difference in the true proportions $p_1 - p_2$ is

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

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution.

Exercise

3 Inference on Two Population Proportions

Consider the process of manufacturing crankshaft bearings described in Example 8.7. Suppose that a modification is made in the surface finishing process and that, subsequently, a second random sample of 85 bearings is obtained. The number of defective bearings in this second sample is 8. Therefore, because $n_1 = 85$, $\hat{p}_1 = 10/85 = 0.1176$, $n_2 = 85$, and $\hat{p}_2 = 8/85 = 0.0941$, we can obtain an approximate 95% confidence interval on the difference in the proportion of defective bearings produced under the two processes from Equation 10.41 as follows:

Consider the process of manufacturing crankshaft bearings described in Example 8.7. Suppose that a modification is made in the surface finishing process and that, subsequently, a second random sample of 85 bearings is obtained. The number of defective bearings in this second sample is 8. Therefore, because $n_1 = 85$, $\hat{p}_1 = 10/85 = 0.1176$, $n_2 = 85$, and $\hat{p}_2 = 8/85 = 0.0941$, we can obtain an approximate 95% confidence interval on the difference in the proportion of defective bearings produced under the two processes from Equation 10.41 as follows:

Solution

$$\hat{p}_1-\hat{p}_2-z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq p_1-p_2 \leq \hat{p}_1-\hat{p}_2+z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\Longrightarrow -0.0685 \leq p_1 - p_2 \leq 0.1155$$

Q&A

Thank you for listening!