Chapter5
Inferences Based on Two Samples

PHOK Ponna and PHAUK Sokkhey

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia

16/11/2023

Contents

1. Z Tests for a Difference Between Two Population Means
2. The Two-Sample t Test and Confidence Interval
3. Analysis of Paired Data

Contents

1 Z Tests for a Difference Between Two Population Means

2 The Two-Sample t Test and Confidence Interval

3 Analysis of Paired Data

Basic Assumptions

1. $X_1, X_2, ..., X_m$ is a random sample from a distribution with mean μ_1 and variance σ_1^2.
2. $Y_1, Y_2, ..., Y_n$ is a random sample from a distribution with mean μ_2 and variance σ_2^2.
3. The X and Y samples are independent of one another.

Theorem 1

The expected value of $\bar{X} - \bar{Y}$ is $\mu_1 - \mu_2$, so $\bar{X} - \bar{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. The standard deviation of $\bar{X} - \bar{Y}$ is

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Test Procedures for Normal Populations with Known Variances

We assume here that both population distributions are normal and that the values of both σ_1^2 and σ_2^2 are known.

- Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$
- Test statistic value: $z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$

Alternative Hypothesis	Rejection Region for Level α Test
$H_a : \mu_1 - \mu_2 > \Delta_0$	$R = \{z : z \geq z_\alpha\}$
$H_a : \mu_1 - \mu_2 < \Delta_0$	$R = \{z : z \leq -z_\alpha\}$
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$R = \{z : z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}\}$

Example 1

Analysis of a random sample consisting of $m = 20$ specimens of cold-rolled steel to determine yield strengths resulted in a sample average strength of $\bar{x} = 29.8$ ksi. A second random sample of $n = 25$ two-sided galvanized steel specimens gave a sample average strength of $\bar{y} = 37.4$ ksi. Assuming that the two yield-strength distributions are normal with $\sigma_1 = 4.0$ and $\sigma_2 = 5.0$ (suggested by a graph in the article "Zinc-Coated Sheet Steel: An Overview," Automotive Engr., Dec. 1984: 39–43), does the data indicate that the corresponding true average yield strengths μ_1 and μ_2 are different? Let's carry out a test at significance level $\alpha = 0.01$.

Alternative Hypothesis	$\beta(\Delta')=$ P(type II error when $\mu_1 - \mu_2 = \Delta'$)
$H_a : \mu_1 - \mu_2 > \Delta_0$	$\Phi\left(z_\alpha - \frac{\Delta' - \Delta_0}{\sigma}\right)$
$H_a : \mu_1 - \mu_2 < \Delta_0$	$1 - \Phi\left(-z_\alpha - \frac{\Delta' - \Delta_0}{\sigma}\right)$
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$\Phi\left(z_{\alpha/2} - \frac{\Delta' - \Delta_0}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\Delta' - \Delta_0}{\sigma}\right)$

where $\sigma = \sigma_{\bar{X}-\bar{Y}} = \sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)}$ Sample sizes m and n can be determined that will satisfy both $P(\text{type I error}) = \alpha$ and $P(\text{type II error when } \mu_1 - \mu_2 = \Delta') = \beta$. For an upper-tailed test, equating the previous expression for $\beta(\Delta')$ to the specified value of β gives

$$\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} = \frac{(\Delta' - \Delta_0)^2}{(z_\alpha + z_\beta)^2}.$$

These expressions are also correct for a lower-tailed test, whereas α is replaced by $\alpha/2$ for a two-tailed test.

Example 2

Persons having Raynaud's syndrome are apt to suffer a sudden impairment of blood circulation in fingers and toes. In an experiment to study the extent of this impairment, each subject immersed a forefinger in water and the resulting heat output ($cal/cm^2/min$) was measured. For $m = 10$ subjects with the syndrome, the average heat output was $\bar{x} = 64$, and for $n = 10$ nonsufferers, the average output was 2.05. Let μ_1 and μ_2 denote the true average heat outputs for the two types of subjects. Assume that the two distributions of heat output are normal with $\sigma_1 = 0.2$ and $\sigma_2 = 0.4$.

- **(a)** Carry out the test $H_0 : \mu_1 - \mu_2 = -1$ versus $H_a : \mu_1 - \mu_2 < -1$. Compute the $P-$value.
- **(b)** What is the probability of a type II error when the actual difference between μ_1 and μ_2 is $\mu_1 - \mu_2 = -1.2$?
- **(c)** Assuming that $m = n$, what sample sizes are required to ensure that $\beta = 0.1$ when $\mu_1 - \mu_2 = -1.2$.

Large-Sample Tests

The assumptions of normal population distributions and known values of σ_1 and σ_2 are unnecessary when both sample sizes are large($m \geq 30$ and $n \geq 30$). In this case, the Central Limit Theorem guarantees that $\bar{X} - \bar{Y}$ has approximately a normal distribution regardless of the underlying population distributions. Furthermore, using S_1^2 and S_2^2 in place of σ_1^2 and σ_2^2 gives a variable whose distribution is approximately standard normal:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

A large-sample test statistic results from replacing $\mu_1 - \mu_2$ by Δ_0, the expected value of $\bar{X} - \bar{Y}$ when H_0 is true.
Use of the test statistic value along with the previously stated upper-, lower-, and two-tailed rejection regions based on z critical values gives large-sample tests whose significance levels are approximately α.

Example 3

Are male college students more easily bored than their female counterparts? This question was examined in the article "Boredom in Young Adults–Gender and Cultural Comparisons" (J. Cross-Cult. Psych., 1991: 209–223). The authors administered a scale called the Boredom Proneness Scale to 97 male and 148 female U.S. college students. Does the accompanying data support the research hypothesis that the mean Boredom Proneness Rating is higher for men than for women? Test the appropriate hypotheses using a 0.05 significance level.

Gender	Sample Size	Sample Mean	Sample SD
Male	97	10.40	4.83
Female	148	9.26	4.68

Confidence Intervals for $\mu_1 - \mu_2$

When both population distributions are normal, standardizing $\bar{X} - \bar{Y}$ gives a random variable Z with a standard normal distribution.
A $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}.$$

If both m and n are large, the CLT implies that this interval is valid even without the assumption of normal populations; in this case, the confidence level is approximately $100(1-\alpha)\%$. Furthermore, use of the sample variances S_1^2 and S_2^2 in the standardized variable Z yields a valid interval in which s_1^2 and s_2^2 replace σ_1^2 and σ_2^2.
Provided that m and n are both large, a CI for $\mu_1 - \mu_2$ with a confidence level of approximately $100(1-\alpha)\%$ is

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}.$$

Example 4

Two brands of batteries are tested, and their voltages are compared. The summary statistics follow. Find the 95% confidence interval of the true difference in the means. Assume that both variables are normally distributed.

Brand X	Brand Y
$\bar{X}_1 = 9.2$ volts	$\bar{X}_2 = 8.8$ volts
$\sigma_1 = 0.3$ volt	$\sigma_2 = 0.1$ volt
$n_1 = 27$	$n_2 = 30$

What does your interval say about the claim that there is no difference in mean voltages?

Contents

1 Z Tests for a Difference Between Two Population Means

2 The Two-Sample t Test and Confidence Interval

3 Analysis of Paired Data

The Two-Sample t-Test

Assumption

Both population distributions are normal, so that $X_1, X_2, ..., X_m$ is a random sample from a normal distribution and so is $Y_1, ..., Y_n$ (with the X's and Y's independent of one another). The plausibility of these assumptions can be judged by constructing a normal probability plot of the x_i's and another of the y_i's.

The test statistic and confidence interval formula are based on the same standardized variable developed in Section 1, but the relevant distribution is now t rather than z.

The Two-Sample t-Test and Confidence Interval

Theorem

When the population distribution are both normal and $\sigma_X^2 \neq \sigma_Y^2$, the standardized variable

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \tag{1}$$

has approximately a t distribution with df ν estimated form the data by

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}} = \frac{[(se_1)^2 + (se_2)^2]^2}{\frac{(se_1)^4}{m-1} + \frac{(se_2)^4}{n-1}}$$

where $se_1 = \frac{s_1}{\sqrt{m}}, se_1 = \frac{s_2}{\sqrt{n}}$ (round ν down to the nearest integer).

The two-sample t confidence interval for $\mu_1 - \mu_2$ with confidence level $100(1-\alpha)\%$ is then

$$\bar{x} - \bar{y} \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}.$$

The two-sample t test for testing $H_0 : \mu_1 - \mu_2 = \Delta_0$ is as follows:

$$\text{Test statistic value: } t = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}}$$

Alternative Hypothesis	Rejection Region for Level α Test
$H_a : \mu_1 - \mu_2 > \Delta_0$	$R = \{t : t \geq t_{\alpha,\nu}\}$
$H_a : \mu_1 - \mu_2 < \Delta_0$	$R = \{t : t \leq -t_{\alpha,\nu}\}$
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$R = \{t : t \leq -t_{\alpha/2,\nu}$ or $t \geq t_{\alpha/2,\nu}\}$

Example 5

A researcher wishes to see if the average weights of newborn male infants are different from the average weights of newborn female infants. She selects a random sample of 10 male infants and finds the mean weight is 7 pounds 11 ounces and the standard deviation of the sample is 8 ounces. She selects a random sample of 8 female infants and finds that the mean weight is 7 pounds 4 ounces and the standard deviation of the sample is 5 ounces. Can it be concluded at $\alpha = 0.05$ that the mean weight of the males is different from the mean weight of the females? Assume that the variables are normally distributed.

Example 6

Find the 95% confidence interval for the true difference in means for the data in Example 5.

Pooled t Procedures

Alternatives to the two-sample t procedures just described result from assuming not only that the two population distributions are normal but also that they have equal variances ($\sigma_1^2 = \sigma_2^2$). That is, the two population distribution curves are assumed normal with equal spreads, the only possible difference between them being where they are centered. The standardized variable

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

has a t distribution with df $\nu = m + n - 2$ and

$$S_p = \sqrt{\left(\frac{m-1}{m+n-2}\right) S_1^2 + \left(\frac{n-1}{m+n-2}\right) S_2^2}$$

S_p is called the **pooled estimator of** σ.

Example 7

Consider the pooled T variable

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

which has a t distribution with $m + n - 2$ df when both population distributions are normal with $\sigma_1 = \sigma_2$.

(a) Use this T variable to obtain a pooled t confidence interval formula for $\mu_1 - \mu_2$.

(b) A sample of ultrasonic humidifiers of one particular brand was selected for which the observations on maximum output of moisture (oz) in a controlled chamber were 14.0, 14.3, 12.2, and 15.1. A sample of the second brand gave output values 12.1, 13.6, 11.9, and 11.2 ("Multiple Comparisons of Means Using Simultaneous Confidence Intervals," J. Qual. Techn., 1989: 232–41).

Use the pooled t formula from part (a) to estimate the difference between true average outputs for the two brands with a 95% confidence interval.

(c) Estimate the difference between the two μ's using the two-sample t interval discussed in this section, and compare it to the interval of part (b).

(d) Carry out the pooled t test $H_0 : \mu_1 - \mu_2 = 0$ versus $H_a : \mu_1 - \mu_2 \neq 0$ at $\alpha = 0.05$.

Contents

1. Z Tests for a Difference Between Two Population Means

2. The Two-Sample t Test and Confidence Interval

3. **Analysis of Paired Data**

Analysis of Paired Data

Assumpions

- $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \ldots, Y_n$ is a random sample from $N(\mu_2, \sigma_2^2)$
- The X and Y samples are dependent.

The data consists of n independently selected pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $E(X_i) = \mu_1$ and $E(Y_i) = \mu_2$. Let $D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n$ so the D_i's are the differences within pairs. Then the D_i's are assumed to be normally distributed with mean value μ_D and variance σ_D^2 (this is usually a consequence of the X_i's and Y_i's themselves being normally distributed).

The Paired t Test

Because different pairs are independent, the D_i's are independent of each other. If we let $D = X - Y$, where X and Y are the first and second observations, respectively, within an arbitrary pair, then the expected difference is $\mu_D = E(X - Y) = E(X) - E(Y) = \mu_1 - \mu_2$. Thus any hypothesis about $\mu_1 - \mu_2$ can be phrased as a hypothesis about the mean difference μ_D. But since the D_i's constitute a normal random sample (of differences) with mean μ_D, hypotheses about μ_D can be tested using a one-sample t test.

Null Hypothesis: $H_0 : \mu_D = \Delta_0$

Test statistics value: $t = \frac{\bar{d} - \Delta_0}{s_D/\sqrt{n}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
$H_a : \mu_D > \Delta_0$	$t \geq t_{\alpha,n-1}$
$H_a : \mu_D < \Delta_0$	$t \leq -t_{\alpha,n-1}$
$H_a : \mu_D \neq \Delta_0$	either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$

A P-value can be calculated as was done for earlier t tests.

Example 8

Bank Deposits A random sample of nine local banks shows their deposits (in billions of dollars) 3 years ago and their deposits (in billions of dollars) today. At $\alpha = 0.05$, can it be concluded that the average in deposits for the banks is greater today than it was 3 years ago? Use $\alpha = 0.05$. Assume the variable is normally distributed.

Bank	1	2	3	4	5	6	7	8	9
3 years ago	11.42	8.41	3.98	7.37	2.28	1.10	1.00	0.9	1.35
Today	16.69	9.44	6.53	5.58	2.92	1.88	1.78	1.5	1.22

Example 9

A dietitian wishes to see if a person's cholesterol level will change if the diet is supplemented by a certain mineral. Six randomly selected subjects were pretested, and then they took the mineral supplement for a 6-week period. The results are shown in the table. (Cholesterol level is measured in milligrams per deciliter.) Can it be concluded that the cholesterol level has been changed at $\alpha = 0.10$? Assume the variable is approximately normally distributed.

Subject	1	2	3	4	5	6
Before	210	235	208	190	172	244
After	190	170	210	188	173	228

Confidence Interval for the Mean Difference

The paired t CI for μ_D is

$$\bar{d} - t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}} \leq \mu_D \leq \bar{d} + t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}$$

Example 10

Find the 90% confidence interval for μ_D for the data in Example 9.