The Difference of Two Sample Means with Unknown Variance

Prerequisites

You should be familiar with (1) determining a confidence interval for a sample mean drawn from a population of known or unknown variance; (2) the biased and unbiased estimator of the population variance.

Unbiased estimator of the population mean

The sample mean, $\overline{X}$, is an unbiased and consistent estimator of the population mean, μ.

Unbiased estimator of the population variance

$s^2 = \frac{n}{n-1}S^2$ is an unbiased estimator of the population variance where $S^2 = \frac{\sum X^2}{n} - \left(\overline{X}\right)^2$ is the sample variance.

The difference of two sample means

Suppose we have two independent samples $(X_1 \text{ and } X_2)$ drawn one from each of two normally distributed populations of differing known variance. To say that the samples are independent is to say that values of X_1 are not in any way associated or paired with values of X_2. That is, we are given that

$X_1 \sim N\left(\mu_1, {\sigma_1}^2\right)$

$X_2 \sim N\left(\mu_2, {\sigma_2}^2\right)$

where σ_1^2 and σ_2^2 are known. Alternatively, if the variances are unknown we shall estimate them using the unbiased estimators to obtain estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. Then the *difference of the two sample means*

$\bar{D} = \bar{X}_1 - \bar{X}_2$

is the linear combination of two normally distributed variables, and hence is itself normally distributed. The distribution is as follows.

$$\bar{D} = \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

where n_1 and n_2 are the sizes of the independent samples of $\bar{X}_1$ and $\bar{X}_2$ respectively. This result enables us to find a confidence interval for the difference of two sample means. The expression

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

is called the *standard error of the difference of two means.* When the population variances are not known, they must be estimated using the unbiased estimator; that is, we substitute the estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ for σ_1^2 and σ_2^2 respectively in the above. Then, strictly, since the variances are estimated, the distribution of the standard error of the difference of the two means $\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$ is **not** normally distributed. However, if the sample sizes are large $\left(\text{both } n_1 \text{ and } n_2 > 30\right)$ then it is approximately normally distributed and the approximation is very close. In that case we may use the normal distribution as before to determine the critical values in the confidence interval.

Example (1)

At a DIY store two different varieties, A and B, of sheet hardboard are sold. The management decided to measure the area, x square metres, of each of 100 randomly selected sheets of variety A. The results are summarised below.

$\sum x = 302.5 \qquad \sum x^2 = 916.25$

The management also measured the area, y square metres, of each of 120 randomly selected sheets of variety B. The results are summarised below.

$\sum y = 388.4 \qquad \sum y^2 = 1258.90$

Determine an approximate 95% confidence interval for the difference in the population means of the areas of the two varieties.

Solution

$\sum x = 302.5 \qquad \sum x^2 = 916.25 \qquad n = 100$

$$\bar{x} = \frac{\sum x}{n} = \frac{302.5}{100} = 3.025$$

$$S_x^2 = \frac{\sum x^2}{n} - \left(\bar{x}\right)^2 = \frac{916.25}{100} - \left(3.025\right)^2 = 0.011875$$

$$\hat{\sigma}_x^{\,2} = s_x^{\,2} = \frac{n}{n-1} S_x^{\,2} = \frac{100}{99} \times 0.011875 = 0.0119949...$$

$$\sum y = 388.4 \qquad \sum y^2 = 1258.90$$

$$\bar{y} = \frac{\sum y}{n} = \frac{388.4}{120} = 3.2366...$$

$$S_y^{\,2} = \frac{\sum y^2}{n} - (\bar{y})^2 = \frac{1258.90}{120} - (3.2366...)^2 = 0.014822...$$

$$\hat{\sigma}_y^{\,2} = s_y^{\,2} = \frac{n}{n-1} S_y^{\,2} = \frac{120}{119} \times 0.014822... = 0.0149467...$$

The estimate for the mean difference of the areas of the two varieties is

$$d = \bar{y} - \bar{x} = 3.2366... - 3.025 = 0.21166...$$

The standard error of the mean difference is

$$\text{SE} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{0.0119949...}{100} + \frac{0.0149467...}{120}} = \sqrt{2.445... \times 10^{-4}} = 0.015636...$$

A 95% confidence interval for the true mean difference of the areas of the two varieties is

$$\bar{d} - 1.96 \times \text{SE} < \mu < \bar{d} + 1.96 \times \text{SE}$$
$$0.21166... - 1.96 \times 0.015636... < \mu < 0.21166... + 1.96 \times 0.015636...$$
$$0.1810... < \mu < 0.2423...$$
$$0.18 < \mu < 0.24 \quad (2 \text{ s.f.})$$

Hypothesis testing

We may also be asked to test whether the two population means are equal at a given significance level.

H_0 $\quad \mu_1 = \mu_2$

H_1 $\quad \mu_1 \neq \mu_2$ $\quad$ (two tailed)

alternatively

H_1 $\quad \mu_1 < \mu_2$ or $\mu_1 > \mu_2$ $\quad$ (one tailed)

Since the difference of the two sample means is

$$\bar{D} = \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

the test statistic is

$$Z_{test} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The critical value is found using the tables for the standardised normal distribution and finding the z value corresponding to the given significance level. We may compare this test statistic with

the critical value for the test depending on the significance level, or alternatively find the probability (*p*-value) that corresponds to the test statistic. The first example illustrates the case where the population variances are known and different.

Example (2)

A random sample is taken from each of two normally distributed populations. The sample size of the first population is 10 with sample mean 18.2. The sample size of the second population is 12, with sample mean 15.3. It is known that the population variances are 2.5 and 2.8 respectively. Test at the 1% level whether the means of the two populations are equal.

Solution

First sample

$X_1 \sim N\left(\mu_1, \sigma_1^2\right) \quad n_1 = 10,\ \bar{X}_1 = 18.2,\ \sigma_1^2 = 2.5$

Second sample

$X_2 \sim N\left(\mu_2, \sigma_2^2\right) \quad n_2 = 12,\ \bar{X}_2 = 15.3,\ \sigma_2^2 = 2.8$

The null and alternative hypotheses are formulated as follows.

$H_0 \quad \mu_1 = \mu_2$

$H_1 \quad \mu_1 \neq \mu_2$ two-tailed, significance level, $\alpha = 1\%$

The test statistic is

$$Z_{test} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{|18.2 - 15.3|}{\sqrt{\frac{2.5}{10} + \frac{2.8}{12}}}$$

$$= 4.17133...$$

$$= 4.171 \ (3 \text{ d.p.})$$

The critical *z*-value is

$Z_{critical} = P(Z > 0.995)$ since α =1% and the test is 2-tailed

$= 2.576$

$Z_{test} > Z_{critical}$

$\therefore$ Accept H_1, reject H_0

The two population means are not equal.

Alternatively, since $P\left(Z < 3.090\right) = 0.999$, the *p*-value (2 tailed) associated with a test value of $z_{test} = 4.171$ is $p < 0.0005$

In this next example the population variances are unknown and must be estimated.

Example (3)

Two random variables X and Y are normally distributed. Independently, samples of both were taken as follows.

$n_1 = 120 \qquad \sum x = 250.2 \qquad \sum x^2 = 1008.4$

$n_2 = 150 \qquad \sum y = 342.8 \qquad \sum y^2 = 1278.5$

Test the hypothesis that the mean of X is less than the mean of Y at the 1% significance level.

Solution

In this question the variances are unknown and therefore must be estimated using the unbiased estimator. Since the sample sizes are large we may use the z-score to evaluate the hypothesis.

$X \sim N\left(\mu_1, \sigma_1^2\right) \qquad n_x = 120 \qquad \sum x = 250.2 \qquad \sum x^2 = 1008.4$

$$\bar{x} = \frac{\sum x}{n} = \frac{250.2}{120} = 2.085$$

$$S_x^{\,2} = \frac{\sum x^2}{n} - \left(\bar{x}\right)^2 = \frac{1008.4}{120} - \left(2.085\right)^2 = 4.0561...$$

$$\hat{\sigma}_x^{\,2} = \frac{n}{n-1} S_x^{\,2} = \frac{120}{119} \times 4.0561... = 4.0901...$$

$Y \sim N\left(\mu_2, \sigma_2^2\right) \qquad n_y = 150 \qquad \sum y = 342.8 \qquad \sum y^2 = 1278.5$

$$\bar{y} = \frac{\sum y}{n} = \frac{342.8}{150} = 2.2853...$$

$$S_y^{\,2} = \frac{\sum y^2}{n} - \left(\bar{y}\right)^2 = \frac{1278.5}{150} - \left(2.2853\right)^2 = 3.30058...$$

$$\hat{\sigma}_x^{\,2} = \frac{n}{n-1} S_x^{\,2} = \frac{150}{149} \times 3.30058... = 3.3227...$$

$H_0 \qquad \mu_1 = \mu_2$

$H_1 \qquad \mu_1 < \mu_2 \qquad$ one-tailed, significance level, $\alpha = 1\%$

$$Z_{test} = \frac{\left|\bar{X} - \bar{Y}\right|}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

$$= \frac{\left|2.085 - 2.2853...\right|}{\sqrt{\frac{4.0901...}{120} + \frac{3.3227...}{150}}}$$

$$= 0.84478...$$

$$= 0.845 \quad (3 \text{ d.p.})$$

$Z_{critical} = P(Z > 0.995) = 2.576 \qquad$ since $\alpha = 1\%$ and the test is 2-tailed

$Z_{\text{test}} < Z_{\text{critical}}$

$\therefore$ Reject H_1, accept H_0

The two population means are equal.

Alternatively, the p-value (1 tailed) associated with a test value of $z_{test} = 0.845$ is $p + 1 - 0.6009 = 0.3991$ which is much greater than the critical value of 0.01. Therefore, as above reject H_1, accept H_0.

When the population variances are identical

When it is given that the population variances are identical then the test is simpler, since then

$$Z_{\text{test}} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where σ^2 is the common variance.

Example (4)

Two independent random samples were taken, one from each of two normally distributed populations both of which may be assumed to have the same variance, σ^2=16. If the first sample had size 8 and mean 4 and the second have size 24 and mean 5, test the null hypothesis $\mu_1 - \mu_2 = 0$ against the alternative hypothesis $\mu_1 - \mu_2 \neq 0$ at the 1% significance level, where μ_1 and μ_2 are the means of the respective populations.

Solution

$X_1 \sim N\left(\mu_1, \sigma_1^2\right) \quad n_1 = 8 \quad \bar{X}_1 = 4$

$X_2 \sim N\left(\mu_2, \sigma_2^2\right) \quad n_2 = 24 \quad \bar{X}_2 = 5$

$\sigma^2 = 16 \qquad \sigma = \sqrt{16} = 4$

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

$H_1 \qquad \mu_1 - \mu_2 \neq 0 \qquad$ two-tailed $\qquad \alpha = 1\%$

$$Z_{test} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{|4 - 5|}{4\sqrt{\frac{1}{8} + \frac{1}{24}}} = \frac{1}{4\sqrt{\frac{3+1}{24}}} = \frac{1}{4\sqrt{\frac{1}{6}}} = \frac{\sqrt{6}}{4} = 0.612 \text{ (3 s.f.)}$$

$Z_{critical} = P(Z < 0.995) = 2.576$

$Z_{test} < Z_{critical}$

$\therefore$ Accept H_0, reject H_1. The two means are the same.