§26. Two Sample Testing and Confidence Intervals

Two samples are drawn from two normal populations. What is the difference of population means?

Case 1: Variances Known

Let $X_{11}, \ldots, X_{1n}-$ sample from $N(\mu_1, \sigma_1)$ and $X_{21}, \ldots, X_{2n}-$ sample from $N(\mu_2, \sigma_2)$

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

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

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

The test statistic when testing:

$H_0 : \mu_1 - \mu_2 = \Delta$ versus $H_1 : \mu_1 - \mu_2 > \Delta$, $\quad H_1 : \mu_1 - \mu_2 < \Delta$, or $H_1 : \mu_1 - \mu_2 \neq \Delta$ is:

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \sim \quad N(0,1)$$

Case 2: Variances unknown but assumed equal

Use the pooled estimator for σ^2

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

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

$$\overline{x}_1 - \overline{x}_2 - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \overline{x}_1 - \overline{x}_2 + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad ; \quad (n_1 + n_2 - 2) \text{ df}$$

The test statistic when testing $H_0 : \mu_1 - \mu_2 = \Delta$ versus
$H_1 : \mu_1 - \mu_2 > \Delta$, $H_1 : \mu_1 - \mu_2 < \Delta$, or $H_1 : \mu_1 - \mu_2 \neq \Delta$ is:

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \sim \quad \text{Student}\left(n_1 + n_2 - 2\right)$$

Case 3: Variance Unknown and $\sigma_1^2 \neq \sigma_2^2$

The variable $t_0 = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ is approximately t-distributed with

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

degrees of freedom.

In this case, t_0, is the test statistic for testing $H_0 : \mu_1 - \mu_2 = \Delta$ versus
$H_1 : \mu_1 - \mu_2 > \Delta$, $H_1 : \mu_1 - \mu_2 < \Delta$, or $H_1 : \mu_1 - \mu_2 \neq \Delta$.

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

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

Example 1

Probstat class is given in two sections, one online and on in-person.
Here are the results for the final grade average in the two sections:

$$\begin{aligned} &\text{Online:} \quad n_1 = 33, \quad \bar{x}_1 = 74.6, \quad s_1^2 = 211.14 \\ &\text{In-person:} \quad n_2 = 28, \quad \bar{x}_2 = 81.5, \quad s_2^2 = 240.09 \end{aligned}$$

Assuming that the population grades are normally distributed and have the same variance construct a 95% confidence interval for $\mu_1 - \mu_2$.

Solution

$$s_p^2 = \frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 - n_2 - 2} = \frac{(32)(211.14) + (27)(240.09)}{33 + 28 - 2}$$
$$= 224.39$$

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 74.6 - 81.5 \pm (2.00)\sqrt{224.39}\sqrt{\frac{1}{33} + \frac{1}{28}}$$
$$= -6.9 \pm 7.70$$

$$-14.6 \le \mu_1 - \mu_2 \le 0.8 \qquad \text{with 95\% confidence.}$$

Based on this CI can we claim that the online section has lower average mark with 95% confidence? → No, since zero is in the interval.

Example 2

Arsenic in drinking water.

Urban communities:	$n_1 = 10$,	$\bar{x}_1 = 12.5$,	$s_1 = 7.63$	(ppm)
Rural commuunities:	$n_2 = 10$,	$\bar{x}_2 = 27.5$,	$s_2 = 15.3$	(ppm)

Is there a difference in the concentration of arsenic in drinking water between urban and rural communities?

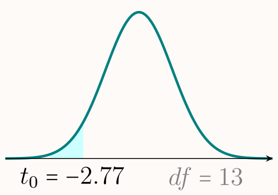
Test $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 < 0$ at $\alpha = 0.05$.

Assume the population concentrations are normally distributed.

Solution

$$t_0 = \frac{12.5 - 27.5}{\sqrt{\frac{(7.63)^2}{10} + \frac{(15.3)^2}{10}}} = -2.77$$

$$\nu = \frac{\left[\frac{(7.63)^2}{10} + \frac{(15.3)^2}{10}\right]^2}{\frac{[(7.63)^2/10]^2}{9} + \frac{[(15.3)/10]^2}{9}} = 13.2 \quad \Rightarrow \quad 13 \text{ df}$$



$$P - value = 0.008 < 0.05 = \alpha$$

$\Rightarrow$ Reject H_0. Accept H_1.

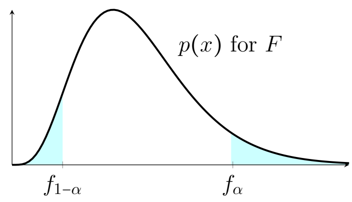
The arsenic concentration in the water of urban communities is lower than in the water of rural communities.

Estimation and Testing for Variances from Two Normal Populations

Definition 1

If $Y_1 \sim \chi^2(k)$ and $Y_2 \sim \chi^2(\ell)$ then

$$F = \frac{Y_1/k}{Y_2/\ell} \sim F(k, \ell)$$



See Kinney for the density

Note: $f_{1-\alpha, k, \ell} = \dfrac{1}{f_{\alpha, \ell, k}}$

Context: Let $x_{11}, \ldots x_{1n_1}$ be a random sample from $N(\mu_1, \sigma_1)$ and let $x_{21}, \ldots, x_{2n_2}$ be a random sample from $N(\mu_2, \sigma_2)$. Let s_1^2 and s_2^2 be the sample variances. Then the ratio statistic is:

$$f_0 = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \quad \text{is} \quad F(n_1 - 1, n_2 - 1)$$

A $100(1-\alpha)\%$ confidence interval on σ_1^2/σ_2^2 is:

$$\frac{s_1^2}{s_2^2} \cdot f_{1-\alpha/2, n_2-1, n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} \cdot f_{\alpha/2, n_2-1, n_1-1}$$

Alternatively using the reciprocal symmetry of the family of F-distributions this same confidence interval could be written as:

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{f_{\alpha/2, n_1-1, n_2-1}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} \cdot f_{\alpha/2, n_2-1, n_1-1}$$

Testing $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 > \sigma_2^2$, $H_1 : \sigma_1^2 < \sigma_2^2$ or $H_1 : \sigma_1^2 \ne \sigma_2^2$ uses the test statistic

$$f_0 = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Example 3

Titanium alloy component for jet turbine engine. Two processes produce the same surface roughness but might have different variances.

$$\begin{aligned} &\text{Process \#1}: \quad n_1 = 11, \quad s_1 = 5.1\,\mu m \\ &\text{Process \#2}: \quad n_2 = 16, \quad s_2 = 4.7\,\mu m \end{aligned}$$

Assuming the populations are normally distributed construct 95% confidence interval for σ_1/σ_2

Solution

$$\frac{s_1^2}{s_2^2}\cdot f_{0.975,\,15,\,10} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot f_{0.025,\,15,\,10}$$

$$\frac{(5.1)^2}{(4.7)^2}\cdot\frac{1}{3.06} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{(5.1)^2}{(4.7)^2}\cdot 3.52$$

$$0.3848 \le \sigma_1^2/\sigma_2^2 \le 4.1446$$

$$0.6203 \le \sigma_1/\sigma_2 \le 2.0358 \qquad \text{with 95\% confidence}$$

Example 4

Two processes for making semiconductor waters achieve same thickness of the oxide layer. Do the oxide layer's thickness have the same variability?

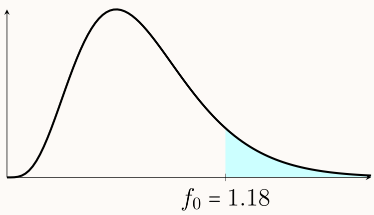
$$\text{Process \#1}: \quad n_1 = 16, \quad s_1 = 2.13\,\text{Å}$$
$$\text{Process \#2}: \quad n_2 = 16, \quad s_2 = 1.96\,\text{Å}$$

Test $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_2 : \sigma_1^2 > \sigma_2^2$ at $\alpha = 0.05$

Solution

$$f_0 = \frac{(2.13)^2}{(1.96)^2} = 1.18$$

$$P - value = 0.3764$$



Fail to reject H_0. This sample does not provide enough evidence to claim that the two processes produce different thickness variabilities.