Tests of differences between two populations

To test if two populations $x$ and $y$ are different we can compare specified distributional properties $a_x$ and $a_y$ (means, variances, 90 percentiles, etc.).

Null hypothesis:

$$H_0: a_x = a_y = a_0 \text{ or } a_x - a_y = 0$$

The hypothesis test may be based on "natural" (unbiased and consistent) estimators of $a_x$ and $a_y$, derived from the independent random samples $x_1, x_2, \ldots, x_{Nx}$ and $y_1, y_2, \ldots, y_{Ny}$:

$$\hat{a}_x = \hat{a}_x(x_1, x_2, \ldots, x_{Nx})$$

$$\hat{a}_y = \hat{a}_y(x_1, x_2, \ldots, x_{Ny})$$

We can derive a two-sided rejection region $R_{a_0}$ written in terms of a standardized statistic $z$, following the same basic procedure as in the single population case (see Class 15):

$$z(\hat{a}_x, \hat{a}_y, a_x, a_y) = \frac{(\hat{a}_x - \hat{a}_y) - (a_x - a_y)}{SD[\hat{a}_x - \hat{a}_y]}$$

$$z(\hat{a}_x, \hat{a}_y, a_0, a_0) = \frac{(\hat{a}_x - \hat{a}_y)}{SD[\hat{a}_x - \hat{a}_y]}$$

$$R_{z0} : z(\hat{a}_x, \hat{a}_y, a_0, a_0) \leq z_L = F^{-1}_z\left(\frac{\alpha}{2}\right)$$

$$z(\hat{a}_x, \hat{a}_y, a_0, a_0) \geq z_U = F^{-1}_z\left(1 - \frac{\alpha}{2}\right)$$

For large samples $z(\hat{a}_x, \hat{a}_y, a_0, a_0)$ has a unit normal distribution if $H_0$ is true ($a_x = a_y = a_0$). Use norminv to compute $z_L$ and $z_U$ from $\alpha$.

We can also define a rejection region $R_{a_0}$ written in terms of the nonstandardized estimates:
\[ R_{a0} : \hat{a}_x - \hat{a}_y \leq \Delta a_L = F_{z}^{-1}\left(\frac{\alpha}{2}\right) SD[\hat{a}_x - \hat{a}_y] \]

\[ \hat{a}_x - \hat{a}_y \geq \Delta a_U = F_{z}^{-1}\left(1 - \frac{\alpha}{2}\right) SD[\hat{a}_x - \hat{a}_y] \]

The two-sided \( p \)-value is obtained from:

\[
1 - p/2 = F_{z}\left[\left(\frac{\hat{a}_x - \hat{a}_y}{SD(\hat{a})}\right) \geq 0 \right]
\]

\[
p/2 = F_{z}\left[\left(\frac{\hat{a}_x - \hat{a}_y}{SD(\hat{a})}\right) \leq 0 \right]
\]

For large samples use \texttt{normcdf} to compute \( p \) from \( \frac{\hat{a}_x - \hat{a}_y}{SD(\hat{a})} \).

Special Case: Large sample test of the difference between two means

If the property of interest is the mean then:

\[ H_0 : a_x = E[x] = a_y = E[y] \text{ or } E[x] - E[y] = 0 \]

Natural estimator of \( E[x] - E[y] \) is \( m_x - m_y \).

In large sample case \( m_x - m_y \) is normal with mean and variance:

\[ E[m_x - m_y] = E[x] - E[y] \quad \text{(unbiased)} \]

\[ Var[(m_x - m_y)] = Var[m_x] + Var[m_y] = \frac{\sigma_x^2}{N_X} + \frac{\sigma_y^2}{N_Y} \quad \text{(consistent)} \]

Construct a large sample test statistic \( z \sim N(0,1) \):

\[ z = \frac{m_x - m_y}{\sqrt{\frac{\sigma_x^2}{N_X} + \frac{\sigma_y^2}{N_Y}}} \approx \frac{m_x - m_y}{\sqrt{\frac{s_x^2}{N_X} + \frac{s_y^2}{N_Y}}} \]

Two-sided rejection region written in terms of \( m_x \) and \( m_y \):
\[ R_{a0}: m_x - m_y \leq \Delta a_L = F^{-1}_z\left(\frac{\alpha}{2}\right)SD[m_x - m_y] \]

\[ m_x - m_y \geq \Delta a_U = F^{-1}_z\left(1 - \frac{\alpha}{2}\right)SD[m_x - m_y] \]

The two-sided p-value is obtained from:

\[ 1 - p/2 = F_z(z) = F_z\left[(m_x - m_y)\left(\frac{s_x^2 + s_y^2}{N_x + N_y}\right)^{-1/2}\right] \quad m_x \geq m_y \]

\[ p/2 = F_z(z) = F_z\left[(m_x - m_y)\left(\frac{s_x^2 + s_y^2}{N_x + N_y}\right)^{-1/2}\right] \quad m_x \leq m_y \]

Example: Comparing crop yields with and without fertilizer application

Consider two agricultural fields, one that is fertilized and one that is not. Yield samples (kg/ha) from the two fields are as follows:

Fertilized \((x)\): 66  41  77  80  52  98  99  74  81  78

Not fertilized \((y)\): 65  88  55  124  66  72  96  71

Test the hypothesis \(H0\): Mean yields are the same with and without fertilizer

\[ m_x = \quad s_x = \quad N_x = \]

\[ m_y = \quad s_y = \quad N_y = \]

\[ z = \quad p = \]

MIT  Copyright 2003 Massachusetts Institute of Technology
Last modified Oct. 8, 2003

3