# 1.017/1.010 Class 17
# Testing Hypotheses about Two Populations

## Tests of differences between two populations

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

Null hypothesis:

$$H0: \; a_x = a_y = a_0 \;\text{ or }\; a_x - a_y = 0$$

The hypothesis test may be based on "natural" (unbiased and consistent) estimators of $a_x$ and $a_y$, derived from the **independent** random samples $\boldsymbol{x}_1$, $\boldsymbol{x}_2,\ldots, \boldsymbol{x}_{Nx}$ and $\boldsymbol{y}_1, \boldsymbol{y}_2,\ldots,\boldsymbol{y}_{Ny}$ :

$$\hat{\boldsymbol{a}}_x = \hat{\boldsymbol{a}}_x(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{Nx})$$

$$\hat{\boldsymbol{a}}_y = \hat{\boldsymbol{a}}_y(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{Ny})$$

We can derive a two-sided rejection region $R_{a0}$ written in terms of a **standardized** statistic $\boldsymbol{z}$ , following the same basic procedure as in the single population case (see Class 15):

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

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

$$R_{z0} : \; \boldsymbol{z}(\hat{\boldsymbol{a}}_x, \hat{\boldsymbol{a}}_y, a_0, a_0) \le z_L = F_{\boldsymbol{z}}^{-1}(\frac{\alpha}{2})$$

$$\boldsymbol{z}(\hat{\boldsymbol{a}}_x, \hat{\boldsymbol{a}}_y, a_0, a_0) \ge z_U = F_{\boldsymbol{z}}^{-1}(1 - \frac{\alpha}{2})$$

For **large samples** $\boldsymbol{z}(\hat{\boldsymbol{a}}_x, \hat{\boldsymbol{a}}_y, a_0, a_0)$ has a unit normal distribution if $H0$ 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 $R_{a0}$ written in terms of the **nonstandardized** estimates:

$$R_{a0}: \hat{a}_x - \hat{a}_y \le \Delta a_L = F_{\boldsymbol{z}}^{-1}(\frac{\alpha}{2}) SD[\hat{\boldsymbol{a}}_x - \hat{\boldsymbol{a}}_y]$$

$$\hat{a}_x - \hat{a}_y \ge \Delta a_U = F_{\boldsymbol{z}}^{-1}(1 - \frac{\alpha}{2}) SD[\hat{\boldsymbol{a}}_x - \hat{\boldsymbol{a}}_y]$$

The two-sided $p$-value is obtained from:

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

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

For large samples use `normcdf` to compute $p$ from $\frac{\hat{a}_x - \hat{a}_y}{SD[\hat{\boldsymbol{a}}]}$.

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

If the property of interest is the mean then:

$$H0: \; a_x = E[\boldsymbol{x}] = a_y = E[\boldsymbol{y}] \;\text{ or }\; E[\boldsymbol{x}] - E[\boldsymbol{y}] = 0$$

Natural estimator of $E[\boldsymbol{x}] - E[\boldsymbol{y}]$ is $\boldsymbol{m}_x - \boldsymbol{m}_y$.

In large sample case $\boldsymbol{m}_x - \boldsymbol{m}_y$ is normal with mean and variance:

$E[\boldsymbol{m}_x - \boldsymbol{m}_y] = E[\boldsymbol{x}] - E[\boldsymbol{y}]$ (unbiased)

$$Var[(\boldsymbol{m}_x - \boldsymbol{m}_y) = Var[\boldsymbol{m}_x] + Var[\boldsymbol{m}_y] = \frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y} \quad \text{(consistent)}$$

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

$$\boldsymbol{z} = \frac{\boldsymbol{m}_x - \boldsymbol{m}_y}{\sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}}} \approx \frac{\boldsymbol{m}_x - \boldsymbol{m}_y}{\sqrt{\frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}}}$$

Two-sided rejection region written in terms of $m_x$ and $m_y$:

$$R_{a0} : m_x - m_y \le \Delta a_L = F_z^{-1}(\frac{\alpha}{2}) SD[\boldsymbol{m}_x - \boldsymbol{m}_y]$$

$$m_x - m_y \ge \Delta a_U = F_z^{-1}(1 - \frac{\alpha}{2}) SD[\boldsymbol{m}_x - \boldsymbol{m}_y]$$

The two-sided p-value is obtained from:

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

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

<u>Example: Comparing crop yields with and without fertilizer application</u>

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

$m_x =$ $s_x =$ $N_x =$
$m_y =$ $s_y =$ $N_y =$

$z =$ $p =$