Statistical Tests for Two Sets of Data Drawn From the Same Bivariate Normal Distribution

Bill Menke, December 6, 2017 (after conversation with Martin Stute)

Assume two groups of (X, Y) data: N_A data $(\boldsymbol{X}_A, \boldsymbol{Y}_A)$ in group A and N_B data $(\boldsymbol{X}_B, \boldsymbol{Y}_B)$ in group B. Under the Null Hypothesis, these data are drawn from the same bivariate normal distribution with:

$$\text{mean: } \mathbf{m}^{\text{true}} = \begin{bmatrix} \bar{X} \\ \bar{Y} \end{bmatrix}^{true} \text{ and } \text{ covariance: } \mathbf{C}^{\text{true}} = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}^{true}$$

The mean and covariance can be estimated from the data via:

$$\mathbf{m}^{\text{true}} \approx \mathbf{m}_{\text{A}}^{\text{est}} \approx \mathbf{m}_{\text{B}}^{\text{est}} \text{ and } \mathbf{C}^{\text{true}} \approx \mathbf{C}_{\text{A}}^{\text{est}} \approx \mathbf{C}_{\text{B}}^{\text{est}}$$

$$\mathbf{m}_{\text{A}}^{\text{est}} = \begin{bmatrix} \bar{X}_A \\ \bar{Y}_A \end{bmatrix}^{est} = \begin{bmatrix} mean(\boldsymbol{X}_A) \\ mean(\boldsymbol{Y}_A) \end{bmatrix} \text{ and } \mathbf{C}_{\mathbf{A}}^{\text{est}} = \begin{bmatrix} [std(\boldsymbol{X}_A)]^2 & cov(\boldsymbol{X}_A, \boldsymbol{Y}_A) \\ cov(\boldsymbol{X}_A, \boldsymbol{Y}_A) & [std(\boldsymbol{Y}_A)]^2 \end{bmatrix} \equiv \begin{bmatrix} \sigma_{XA}^2 & \sigma_{XAYA} \\ \sigma_{XAYA} & \sigma_{YA}^2 \end{bmatrix}$$

$$\mathbf{m}_{\text{B}}^{\text{est}} = \begin{bmatrix} \bar{X}_B \\ \bar{Y}_B \end{bmatrix}^{est} = \begin{bmatrix} mean(\boldsymbol{X}_B) \\ mean(\boldsymbol{Y}_B) \end{bmatrix} \text{ and } \mathbf{C}_{\mathbf{A}}^{\text{est}} = \begin{bmatrix} [std(\boldsymbol{X}_B)]^2 & cov(\boldsymbol{X}_B, \boldsymbol{Y}_B) \\ cov(\boldsymbol{X}_B, \boldsymbol{Y}_B) & [std(\boldsymbol{Y}_B)]^2 \end{bmatrix} \equiv \begin{bmatrix} \sigma_{XB}^2 & \sigma_{XBYB} \\ \sigma_{XBYB} & \sigma_{YB}^2 \end{bmatrix}$$

Note that Matlab has mean(), std() and cov() functions. The covariance of the sample mean is a factor of $1/N$ smaller than the covariance of the data:

$$\mathbf{C}_{\mathbf{mA}}^{\text{est}} = \frac{1}{N_A} \begin{bmatrix} \sigma_{XA}^2 & \sigma_{XAYA} \\ \sigma_{XAYA} & \sigma_{YA}^2 \end{bmatrix} \text{ and } \mathbf{C}_{\mathbf{mB}}^{\text{est}} = \frac{1}{N_B} \begin{bmatrix} \sigma_{XB}^2 & \sigma_{XBYB} \\ \sigma_{XBYB} & \sigma_{YB}^2 \end{bmatrix}$$

Now form a vector of the estimated means $[\bar{X}_A, \bar{Y}_A, \bar{X}_B, \bar{Y}_B]^{estT}$. It has covariance:

$$\begin{bmatrix} \sigma_{XA}^2/N_A & \sigma_{XAYA}/N_A & 0 & 0 \\ \sigma_{XAYA}/N_A & \sigma_{YA}^2/N_A & 0 & 0 \\ 0 & 0 & \sigma_{XB}^2/N_B & \sigma_{XBYB}/N_B \\ 0 & 0 & \sigma_{XBYB}/N_B & \sigma_{YB}^2/N_B \end{bmatrix}$$

The zeros arise because the A and B groups of measurements are independent. The vector $\mathbf{z} = [\Delta\bar{X}, \Delta\bar{Y}]^{\mathbf{T}}$ of the differences between means is formed as

$$\mathbf{z} = \begin{bmatrix} \Delta X \\ \Delta Y \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} X_A \\ Y_A \\ X_B \\ Y_B \end{bmatrix} = \begin{bmatrix} \bar{X}_A - \bar{X}_B \\ \bar{Y}_A - \bar{Y}_B \end{bmatrix}$$

The variance of differences is compute by standard error propagation:

$$\mathbf{C}_{\mathbf{z}} = \begin{bmatrix} \sigma_{\Delta X}^2 & \sigma_{\Delta X \Delta Y} \\ \sigma_{\Delta X \Delta Y} & \sigma_{\Delta Y}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{XA}^2/N_A & \sigma_{XAYA}/N_A & 0 & 0 \\ \sigma_{XAYA}/N_A & \sigma_{YA}^2/N_A & 0 & 0 \\ 0 & 0 & \sigma_{XB}^2/N_B & \sigma_{XBYB}/N_B \\ 0 & 0 & \sigma_{XBYB}/N_B & \sigma_{YB}^2/N_B \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{XA}^2/N_A & \sigma_{XAYA}/N_A \\ \sigma_{XAYA}/N_A & \sigma_{YA}^2/N_A \\ -\sigma_{XB}^2/N_B & \sigma_{XBYB}/N_B \\ \sigma_{XBYB}/N_B & -\sigma_{YB}^2/N_B \end{bmatrix} =$$

$$\begin{bmatrix} \sigma_{XA}^2/N_A + \sigma_{XB}^2/N_B & \sigma_{XAYA}/N_A - \sigma_{XBYB}/N_B \\ \sigma_{XAYA}/N_A - \sigma_{XBYB}/N_B & \sigma_{YA}^2/N_A + \sigma_{YB}^2/N_B \end{bmatrix}$$

Note that the joint p.d.f. $p(\Delta\bar{X}, \Delta\bar{Y})$ is correlated. The correlation can be removed by transforming to a new variable $\boldsymbol{\gamma} = \mathbf{C}_{\mathbf{z}}^{-1/2}\,\mathbf{z}$, which has unit covariance $\mathbf{C}_{\boldsymbol{\gamma}} = \mathbf{I}$.

We now modify the Null Hypothesis: The data are drawn from the same bivariate normal distribution, so that the any difference between their estimated means is due to random variation and consequently the squared length of $\boldsymbol{\gamma}$ (that is, $\chi^2 = \gamma_1^2 + \gamma_2^2$) differs from zero only because of random variation. The variable χ^2 is chi-squared distributed with 2 degrees of freedom, so a Chi-squared test can be used to assess the probability of the observed χ^2.

We can also address the Null Hypothesis: The data are drawn from the same bivariate normal distribution, so that the any difference between their estimated covariances is due to random variation.

Actually, we examine the estimated scatter matrices, $\mathbf{S}_A$ and $\mathbf{S}_B$, defines as

$$\mathbf{S}_A = (N_A - 1)\begin{bmatrix} \sigma_{XA}^2 & \sigma_{XAYA} \\ \sigma_{XAYA} & \sigma_{YA}^2 \end{bmatrix} \quad \text{and} \quad \mathbf{S}_B = (N_B - 1)\begin{bmatrix} \sigma_{XB}^2 & \sigma_{XBYB} \\ \sigma_{XBYB} & \sigma_{YB}^2 \end{bmatrix}$$

because the scatter matrix is known to be Wishart-distributed with:

$$\bar{\mathbf{S}} = N\begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \quad \text{and} \quad \text{var}(\bar{\mathbf{S}}) = N\begin{bmatrix} 2(\sigma_X^2)^2 & (\sigma_{XY})^2 + \sigma_X^2\sigma_Y^2 \\ (\sigma_{XY})^2 + \sigma_X^2\sigma_Y^2 & 2(\sigma_Y^2)^2 \end{bmatrix}$$

Since the elements of the scatter matrix are sums of random variables, we can invoke the central limit theorem and assert that when N is large, the they will be approximately normally distributed. The difference

$$\Delta\bar{\mathbf{S}} = \bar{\mathbf{S}}_{\mathrm{A}} - \bar{\mathbf{S}}_{\mathrm{B}}$$

will then have an expected value of zero with variance $2\,\text{var}(\bar{\mathbf{S}})$. A Z test can then be used assess the probability p_{11} of element $[\Delta\bar{\mathbf{S}}]_{11}$ under the Null Hypothesis (and similarly for the other elements). The total probability is the product $p_{11}p_{12}p_{22}$. Note, however that we must substitute estimated values of the σ's into the formula for $\text{var}(\bar{\mathbf{S}})$, which is only approximate.