Supporting Information (SI)

Formulae, Program

SI(1) **Mean value**. Mean value of a statistical sample $x_1, x_2, ..., x_n$ is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

SI(2) **Variance.** Variance of a statistical sample $x_1, x_2, ..., x_n$ is

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \ .$$

SI(3) **Unbiased variance (or true variance)**. An unbiased estimator of the population variance is

$$\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \ .$$

SI(4) **Standard deviation**. Standard deviation of a sample is the positive square root of the variance, i.e. s_n or σ_n.

SI(5) **Standard deviation of the sample mean.**

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)}} \ .$$

SI(6) **Binomial distribution.** A random variable *X* is binomially distributed if

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \ .$$

SI(7) **Binomial distribution.**

Expected value of a binomially distributed random variable *X* is $E(X) = np$.

Variance of a binomially distributed random variable *X* is $D^2(X) = np(1-p)$.

SI(8) **Kolmogorov – Smirnov test.** One sample.

The empirical distribution function F_n for observations $x_1, x_2, ..., x_n$ is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 1 \; if \;\; x_i \le x \\ 0 \; otherwise \end{cases}$$

The Kolmogorov-Smirnov test statistics are given by

$$K_n = \sqrt{n} D_n = \sqrt{n}\max\left|F_n(x) - F(x)\right|$$

where $F(x)$ is the hypothesized distribution function. The probability distribution of this statistics, given that the null hypothesis of equality of distributions is true, does not depend on what the hypothesized distribution is, as long as it is continuous. The limit distribution of K_n is known. $\lim_{n\to\infty} P\left(\sqrt{n}D_n < z\right) = K(z)$

SI(9) **Kolmogorov – Smirnov test.** Two sample.

To compare two experimental cumulative distributions $F_n(x)$ containing n events, and $G_m(x)$ containing m events, calculate:

$$D_{n,m} = \max\left|F_n(x) - G_m(x)\right| \qquad \text{Then } K_{n,m} = \sqrt{\frac{nm}{n+m}} D_{n,m}$$

is the test statistic for which the confidence levels are the same as the one sample K - S test.

SI(10) **Wilcoxon rank sum test.**

The Wilcoxon rank sum test is used to determine if two populations have the same probability distribution, or whether one population lies to the right or left the other. If $x_1^* < x_2^* < \cdots < x_n^*$ represent the elements of one population in ascending order and $y_1^* < y_2^* < \cdots < y_m^*$ are the elements of the second, let us denote the union of the two population by $z_1^* < z_2^* < \cdots < z_{k+m}^*$. The rank of $rank\, of\; x_i^* = r_i = j$ if $x_i^* = z_j^*$, that is x_i^* stands in the j^{th} position in the sequence $z_1^* < z_2^* < \cdots < z_{k+m}^*$.

The Wilcoxon statistics is $W_{x,y} = \sum_{i=1}^{n} (r_i - i) = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2}$.

It was proved that $E[W_{x,y}] = \frac{n \cdot m}{2}$ and $D^2[W_{x,y}] = \frac{n \cdot m \cdot (n+m+1)}{12}$.

The limit distribution of $W_{x,y}$ under the hypothesis H_0 that X and Y has the same probability distribution is Gaussian:

$$\lim_{n,m\to\infty} P\left(\frac{W_{x,y} - E[W_{x,y}]}{D[W_{x,y}]} < x \middle| H_0 \right) = N(0,1).$$

SI(11) **Normal (Gaussian) distribution.** A random variable X is normally distributed if its density function is

$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2} \right)$ where $\sigma > 0$. The expected value of aGaussian random variable is $E[X] = m$ and the standard deviation is $D[X] = \sigma$.

SI(12) **Student's t-test. (two sample)** A test of the null hypothesis that the means of two normally distributed populations are equal.

Assumptions: independent samples, normal distribution of data, equality of variances.

$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(n-1)s_1^2 + (m-1)s_2^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$ with degree of freedom n+m-2.

SI(13) **Welch two sample test.** A test of the null hypothesis that the means of two normally distributed populations are equal.

Assumptions: independent samples, normal distribution of data, non-equality of variances.

$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$ with degree of freedom $df = \frac{(n-1)(m-1)}{(n-1)c^2 + (m-1)(1-c^2)}$

where $c^2 = \frac{\frac{s_2^2}{m}}{\frac{s_1^2}{n} + \frac{s_2^2}{m}}$.

SI(14) **Beta distribution.** Distribution of a random variable X is called beta distribution if its density function is

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \text{ where } 0 \le x \le 1.$$

Expected value of the beta distribution is $E[X] = \frac{a}{a+b}$

Variance is $D^2[X] = \frac{ab}{(a+b)^2(a+b+1)}$.

SI(15) **Confidence interval of the variance**

From a sample $x_1, x_2, ..., x_n$ with given sample variance s^2, a 95% confidence interval of the parametric variance (σ^2) is given by:

$$\frac{(n-1)s^2}{{\chi_{0.025}}^2(n-1)} > \sigma^2 > \frac{(n-1)s^2}{{\chi_{0.975}}^2(n-1)}$$ (two-sided), or by $$\sigma^2 < \frac{(n-1)s^2}{{\chi_{0.05}}^2(n-1)}$$ (one-sided).

SI(16) **Central Limit Theorem and Chebyshev Inequality**

Central limit theorem: if $X_1, ..., X_n$ are independent identically distributed random variables with finite mean and variance μ, σ^2, then $(X_1 + ... + X_n - n\mu) \cdot \sigma^{-1} n^{-1/2}$ converges in distribution to the standard normal random variable. In particular $\frac{1}{n}(X_1 + ... + X_n)$ is approximately normally distributed with mean μ and variance $\sigma^2 / n.$.

For a random variable with finite mean and variance E, D^2, the Chebyshev inequality reads $P(|Z - E[Z]| \ge \varepsilon) \le \frac{D^2[Z]}{\varepsilon^2}$. In particular for:

$Z_n = \frac{1}{n}(X_1 + ... + X_n)$, $\mu = 0$, $P(|Z_n| \ge \varepsilon) \le \frac{\sigma^2}{n\varepsilon^2}$, for any $n \in N$.

SI(17) **R language and environment (used for calculations)**

R is a free software environment for statistical computing and graphics. The Comprehensive R Archive Network (CRAN) is available at a several URL's, hosted by Departments of Statistics and/or Mathematics in the main countries (e.g. http://cran.at.r-project.org, hosted by the Department of Statistics and Mathematics of the Wirtschaftsuniversität Wien).