Collection of Statistical Formulas

Frequency and cumulated frequency

indices	$i = 1, \ldots, n$ counts the elements in the original data $j = 1, \ldots, m$ value ordinal number (in increasing order) $k = 1, \ldots, K$ counts the classes
Classification/binning of data values	x_k^u with $x_{k+1}^u > x_k^u$: lower class boundary $x_k^o = x_{k+1}^u$ upper class boundary $x_k^* = \frac{1}{2}(x_k^u + x_k^o)$ class center (sometimes simply x_k) $\Delta x_k = x_k^o - x_k^u$ class width
absolute frequency	h_j or h_k
relative frequency	$f_j = \dfrac{h_j}{n}$ or $f_k = \dfrac{h_k}{n}$
density	$h_k^D = \dfrac{h_k}{\Delta x_k}, \quad f_k^D = \dfrac{f_k}{\Delta x_k}$ (only for binned data!)
absolute cumulative frequency	$H_j = H(X \le x_j) = \sum_{j'=1}^{j} h_{j'}$ (or index k for classes) $H(X > x_j) = n - H(X \le x_j)$
relative cumulative frequency	$F_j = F(X \le x_j) = \sum_{j'=1}^{j} f_{j'} = H_j/n$ $F(X > x_j) = 1 - F(X \le x_j)$
Empirical distribution function (original/unbinned data)	$F(x) = \begin{cases} 0 & \text{if } x < x_1, \\ 1 & \text{if } x > x_m, \\ F_j & \text{if } x_j \le x < x_{j+1} \end{cases}$
Empirical distribution function (binned data)	$F(x) = \begin{cases} 0 & \text{if } x < x_1^u, \\ 1 & \text{if } x \ge x_K^o, \\ F_{k-1} + \left(\frac{x - x_k^u}{\Delta x_k}\right) f_k & \text{if } x_k^u \le x < x_k^o \end{cases}$
Empirical density (binned data only)	$f(x) = \dfrac{\mathrm{d}F(x)}{\mathrm{d}x} = \begin{cases} 0 & \text{if } x < x_1^u \text{ or } x > x_K^o, \\ f_k^D & \text{if } x_k^u \le x < x_k^o \end{cases}$

Location scales)

arithmetic means

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(raw data)}$$
$$= \frac{1}{n}\sum_{j=1}^{m} h_j x_j = \sum_{j=1}^{m} f_j x_j \quad \text{(ordered data)}$$
$$\approx \frac{1}{n}\sum_{k=1}^{K} h_k x_k^* = \sum_{k=1}^{K} f_k x_k^* \quad \text{(binned data)}$$

arithmetic means of the linear combination Y=a+bx

$$\bar{y} = a + b\bar{x}$$

harmonisches means

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \quad \text{or} \quad \frac{1}{\sum_{j=1}^{m} \frac{f_j}{x_j}} \quad \text{or} \quad \frac{1}{\sum_{k=1}^{K} \frac{f_k}{x_k^*}}$$

geometric means

$$\bar{x}_G = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

Mode
(binned data)

$$\bar{x}_M = x_{\hat{k}}^u + \frac{f_{\hat{k}}^D - f_{\hat{k}-1}^D}{2f_{\hat{k}}^D - f_{\hat{k}-1}^D - f_{\hat{k}+1}^D}\Delta x_{\hat{k}}$$

$\hat{k}$: class index of the bin with maximum density

Median (ordered data)

$$x_{0.5} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & (n \text{ uneven}) \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right) & (n \text{ even}) \end{cases}$$

Median (binned data)

$$x_{0.5} = x_{k'}^u + \frac{0.5 - F_{k'-1}}{f_{k'}}\Delta x_{k'}$$

with k' such that $F_{k'-1} < 0.5$ but $F_{k'} \geq 0.5$,
(the class containing the distribution function value $F = 0.5$)

q-quantile

$$x_q = x_{k'}^u + \frac{q - F_{k'-1}}{f_{k'}}\Delta x_{k'}$$

with k' as above but replacing 0.5 with q

Dispersion scales

Note: For binne data (classes $k = 1, \ldots, K$) all equations are valid approximately.

Variance	$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{or} \quad \sum_{k=1}^{K} f_k(x_k^* - \bar{x})^2$ Samples/inductive statistics: additional factor $n/(n-1)$ or $n/(n-p)$ if p parameters are estimated
Variance of the linear combination Y=a+bx	$s_y^2 = b^2 s_x^2$
Standard deviation	$s_x = \sqrt{s_x^2}$
Coefficient of variance	$V = \frac{s_x}{\bar{x}}$ (only usefull if all $x_i > 0$)
Alternative formulation	$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$ or $\sum_{k=1}^{K}(x_k^* - \bar{x})^2 f_k = \sum_{k=1}^{K}(x_k^*)^2 f_k - \bar{x}^2$
Range	$R = x_{\max} - x_{\min}$
mean absolute deviation (MAD)	$s_{\mathrm{MAD}} = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i - x_{0.5} \rvert$
Interquartile distance	$s_{\mathrm{IQ}} = x_{0.75} - x_{0.25}$

Measures for the shape of the distribution

$N^{\mathbf{th}}$ central moment	$M_N = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^N$ (raw data) $M_N = \sum_{j=1}^{m}(x_j - \bar{x})^N f_j$ (ordered list) $M_N = \sum_{k=1}^{K}(x_k^* - \bar{x})^N f_k$ (binned data)
Skew	$\Gamma = \frac{M_3}{s_x^3}$
Kurtosis	$K = \frac{M_4}{s_x^4} - 3$

Measures of concentration

Note: Only useful if the feature sum makes sense.

Feature sum $M = \sum_{i=1}^{n} x_i = n\bar{x} = \sum_{k=1}^{K} x_k^* h_k$ (letzteres für binned data)

percentage of M $p_i = \frac{x_i}{M}$ or $p_k = \frac{x_k^* h_k}{M} = \frac{x_k^* f_k}{\bar{x}}$

cumulative percentage $P_i = \sum_{i'=1}^{i} p_{i'}$ (raw and binned data)

Herfindahl index $K_H = \sum_{i=1}^{n} p_i^2$

Exponential index $K_E = \prod_{i=1}^{n} p_i^{p_i} = p_1^{p_1} p_2^{p_2} \dots p_n^{p_n}$

Points on the Lorenz curve $(0,0)$ and (F_i, P_i), $\quad i = 1, \dots, n$ or $1, \dots, K$
where F_i is the usual cumulated percentage of the data.

Gini coeffivient $G = 1 - \sum_{i=1}^{n} (P_i + P_{i-1}) \frac{1}{n} = 1 - \frac{1}{n} \left(2 \sum_{i=1}^{n-1} P_i + P_n \right)$ (raw data)

$= 1 - \sum_{k=1}^{K} (P_k + P_{k-1}) f_k$ (binned data)

Ratio and index metrics

Wachstumsfaktor $I_t = \frac{x_t}{x_{t-1}}$, **Wachstumsrate** $r_t = I_t - 1$

Preisindex von Laspeyres $P_{0t}^{(\mathrm{L})} = \frac{\sum_{i=1}^{n} p_i(t) q_i(0)}{\sum_{i=1}^{n} p_i(0) q_i(0)}$ mit $p_i(t)$ den Preisen und $q_i(t)$ den Mengen

Preisindex von Paasche $P_{0t}^{(\mathrm{P})} = \frac{\sum_{i=1}^{n} p_i(t) q_i(t)}{\sum_{i=1}^{n} p_i(0) q_i(t)}$

Mengenindices von Laspeyres und Paasche $Q_{0t}^{(\mathrm{L})} = \frac{\sum_{i=1}^{n} p_i(0) q_i(t)}{\sum_{i=1}^{n} p_i(0) q_i(0)}$, $\quad Q_{0t}^{(\mathrm{P})} = \frac{\sum_{i=1}^{n} p_i(t) q_i(t)}{\sum_{i=1}^{n} p_i(t) q_i(0)}$

Wertindex $W_{0t} = \frac{\sum_{i=1}^{n} p_i(t) q_i(t)}{\sum_{i=1}^{n} p_i(0) q_i(0)}$

Bivariate analyse: cross-classified data

Relative frequency	$f(x_i, y_j) = \dfrac{h(x_i, y_j)}{n}$
Absolute marginal sum	$h(x_i) = \sum_{j=1}^{K_y} h(x_i, y_j),\ h(y_j) = \sum_{i=1}^{K_x} h(x_i, y_j)$
Relative marginal sum	$f(x_i) = \dfrac{h(x_i)}{n},\ f(y_j) = \dfrac{h(y_j)}{n}$
Conditional percentage	$f(x_i \vert y_j) = \dfrac{h(x_i, y_j)}{h(y_j)} = \dfrac{f(x_i, y_j)}{f(y_j)}$
empirical independence	$f(x_i, y_j) = f(x_i) f(y_j)$ für alle i, j
Arithmetic means	$\bar{x} = \dfrac{1}{n} \sum_{i=1}^{K_x} x_i^* h(x_i) = \sum_{i=1}^{K_x} x_i^* f(x_i)$ $\bar{y} = \dfrac{1}{n} \sum_{j=1}^{K_y} y_j^* h(y_j) = \sum_{j=1}^{K_y} y_j^* f(y_j)$

Bivariate analysis II: simple regression:

(Empirical) covariance

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \text{ (raw data)}$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{K_x}\sum_{j=1}^{K_y}(x_i^*-\bar{x})(y_j^*-\bar{y})h(x_i,y_j) \text{ (cross-classified data)}$$

Alternative formulation

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y} \quad \text{or}$$

$$\sum_i\sum_j(x_i^*-\bar{x})(y_j^*-\bar{y})h(x_i,y_j) = \sum_i\sum_j x_i^*y_j^*h(x_i,y_j) - n\bar{x}\bar{y}$$

Simple linear regression

$$\hat{y}(x) = a + bx, \quad a = \bar{y} - b\bar{x}, b = \frac{s_{xy}}{s_x^2}$$

Nonlinear regression

Minimize the SSE $S(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}(x,a,b,\ldots))^2$

with respect to $a, b, \ldots \Rightarrow \frac{\partial F}{\partial a} = 0,\ \frac{\partial F}{\partial b} = 0, \ldots$

Sensitivity

$$g(x) = \frac{\mathrm{d}\hat{y}}{\mathrm{d}x}$$

Elasticity function

$$\epsilon_{yx} = \frac{x}{\hat{y}}\frac{\mathrm{d}\hat{y}}{\mathrm{d}x}$$

Split into explained and residual variance

$$(y_i-\bar{y}) = \Delta_i + e_i \text{ mit } \Delta_i = \hat{y}_i - \bar{y}, \quad e_i = y_i - \hat{y}_i$$

Additivity of variance components (linear regression)

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}\Delta_i^2 + \sum_{i=1}^{n}e_i^2$$

Coefficient of determination

$$B = 1 - U = 1 - \frac{\sum_{i=1}^{n}e_i^2}{ns_y^2} = 1 - \frac{s_e^2}{s_y^2} \text{ (general)}$$

$$B = \frac{s_\Delta^2}{s_y^2} = \frac{\sum_{i=1}^{n}\Delta_i^2}{ns_y^2} \text{ (simple linear regression)}$$

Bivariate Analysis III: Correlations

Pearson's correlation coefficient	$r_{xy} = \dfrac{s_{xy}}{s_x s_y}$
Relation to the coefficient of determination of simple linear regression	$B = r_{xy}^2$
Spearman's rank correlation coefficient	$r_s = 1 - \dfrac{6\sum_{i=1}^n (R_i^x - R_i^y)^2}{n(n^2-1)}$

Time series

Additive composition	$Y_i = T_i + K_i + S_i + U_i = G_i + S_i + U_i$ $T =$ trend, $K =$ economic situation, $G = T{+}K =$ smooth component, $S =$ seasonal component(s), $U =$ residual.
multiplicative composition	$Y_i = T_i K_i S_i U_i = G_i S_i U_i$
moving average of order τ	$\bar{y}_i^{(\tau)} = \begin{cases} \dfrac{1}{\tau} \sum\limits_{j=i-m}^{i+m} y_j & \tau \text{ odd}, m = \frac{\tau-1}{2}, \\ \dfrac{1}{\tau}\left(\dfrac{y_{i-m} + y_{i+m}}{2} + \sum\limits_{j=i-m+1}^{i+m-1} y_j \right) & \tau \text{ even}, m = \frac{\tau}{2}. \end{cases}$
(Asymmetric) Exponential average	$\hat{y}_t = \alpha y_t + (1-\alpha)\hat{y}_{t-1},\ \hat{y}_0 = y_0$ smoothing parameter $\alpha = 1 - e^{-\frac{1}{\tau}}$
Calculation of the seasonal component (known period)	$\tilde{S}_j = \dfrac{1}{n}\sum\limits_{i=1}^n (y_{ij} - \bar{y}_{ij}^{(\tau)}), \quad S_j = \tilde{S}_j - \dfrac{1}{\tau}\sum\limits_{j'=1}^{\tau} \tilde{S}_{j'},$ $i =$ cycle index, $j =$ index within a cyle. Without trend and economic changes ($G =$ const), $\tilde{S}_j = S_j$ and $\bar{y}_{ij}^{(\tau)} = \bar{y}$.

Chance and probability

Conditional probability	$P(A\|B) = P(A, \text{ if } B\ )$
Stochastic independence	$P(A\|B) = P(A)$ or $P(B\|A) = P(B)$
Probability of union events	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Probability of intersecting events	$P(A \cap B) = P(B)P(A\|B) = P(A)P(B\|A)$
Total probability of exclusive events A_k	$P(B) = \sum_k P(B\|A_k)P(A_k)$
Bayes' theorem	$P(A_k\|B) = \dfrac{P(B\|A_k)P(A_k)}{P(B)}$

Random variables (RV) and distribution function

Probability function of discrete RVs	$p(x_i) = P(X = x_i) = p_i$
Distribution function of discrete RVs	$F(x) = P(X \leq x) = \sum\limits_{x_i \leq x} p(x_i)$
Density function of continuous RV	$f(x) = \frac{dF}{\mathrm{d}x}$
Distribution function of continuous RVs	$F(x) = P(X \leq x) = \int\limits_{x'=-\infty}^{x} f(x')\ \mathrm{d}x'$
Probability of an intervall (X discrete or continuous)	$P(a \leq X \leq b) = F(b) - F(a)$ $P(X > a) = 1 - F(a)$
Expectation	$\begin{aligned} E(X) &= \sum\limits_i x_i p(x_i) && \text{(discrete RV)} \\ &= \int\limits_{x=-\infty}^{\infty} x f(x)\ \mathrm{d}x && \text{(continuous RV)} \end{aligned}$
Variance	$\begin{aligned} V(X) &= E(X - E(X))^2 \\ &= \sum\limits_i [x_i - E(X)]^2 p(x_i) && \text{(discrete RV)} \\ &= \int\limits_{x=-\infty}^{\infty} [x - E(X)]^2 f(x)\ \mathrm{d}x && \text{(continuous RV)} \end{aligned}$
Covariance	$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$ $\begin{aligned} E(XY) &= \sum\limits_i \sum\limits_j x_i y_j p(x_i, y_j) && \text{(diskrete RV)} \\ &= \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} xy f(x, y)\ \mathrm{d}x\ \ \mathrm{d}y && \text{(continuous RV)} \end{aligned}$

Discrete theoretical distributions

Faculty	$n! = n \cdot (n-1) \cdot (n-2) \cdots (1)$
Binomial coefficient	$\left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N(N-1)\ldots(N-n+1)}{n!} = \frac{N!}{n!(N-n)!}$
Binomial distribution $X \sim B(n;\vartheta)$	$P(X=x) = p_B^{(n,\vartheta)}(x) = \left(\begin{array}{c} n \\ x \end{array} \right) \vartheta^x (1-\vartheta)^{n-x}$
Hypergeometric distribution $X \sim H(N;n;M)$	$P(X=x) = p_H^{(N,n,M)}(x) = \frac{\left(\begin{array}{c} M \\ x \end{array} \right)\left(\begin{array}{c} N-M \\ n-x \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)}$
Poisson distribution $X \sim \mathbf{Po}(\mu)$	$P(X=x) = p_{\mathrm{P}}^{(\mu)}(x) = \frac{\mu^x e^{-\mu}}{x!}$

Continuous theoretical distributions

Density of the uniform distribution $X \sim G(a,b)$	$f_G^{(a,b)}(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{sonst.} \end{cases}$
Exponential distribution $X \sim E(\lambda)$	$f_E^{(\lambda)}(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{sonst,} \end{cases} \quad E(X) = \sqrt{V(X)} = \frac{1}{\lambda}$
Relation between the exponential and Poisson distributions	$n \sim \mathrm{Po}(\mu) \Leftrightarrow \Delta \sim E(\mu/T)$ $n =$ Zahl der Ereignisse im Zeitraum T $\Delta =$ distance between two events
Normal distribution $X \sim N(\mu,\sigma^2)$	$f_N^{(\mu,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ E(X) = \mu,\ V(X) = \sigma^2$
Standard normal distribution	$Z = \frac{X-\mu}{\sigma} \sim N(0;1),\ F(z) = F_N^{(0,1)}(x) =: \Phi(z)$
Normalisation $F_N \to \Phi$	$F_N^{(\mu,\sigma^2)}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ $\Phi(-x) = 1 - \Phi(x)$

Functionen of RV and Central Limit Theorem

Density of a (monotonously increasing) function $Z = g(X)$ of a RV	$f_Z(z) = \left[\frac{f_X(x)}{g'(x)}\right]_{x=g^{-1}(z)}$
Density of the sum $Z = X_1 + X_2$ of two independent RV	$f_Z(z) = \int_{-\infty}^{\infty} f_1(x) f_2(z-x)\mathrm{d}x$ where $f_i(x)$ denote the probability densities of X_i
Distribution function of the sum $Z = X_1 + X_2$ of two independent RV	$F_Z(z) = \int_{-\infty}^{\infty} f_1(x) F_2(z-x)\mathrm{d}x = \int_{-\infty}^{\infty} F_1(x) f_2(z-x)\mathrm{d}x$ where $F_i(x)$ denote the distribution functions of X_i
Special case normal distribution	$X_1 \sim N(\mu_1, \sigma_1^2),\ \ X_2 \sim N(\mu_2, \sigma_2^2) \ \Rightarrow\ Z \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$
Expectation and Variance of $Z = aX + b$	$E(Z) = aE(X) + b, \quad V(Z) = a^2 V(X)$
Expectation and Variance of (possibly correlated) $Z = aX + bY$	$E(Z) = aE(X) + bE(Y)$ $V(Z) = a^2V(X) + b^2V(Y) + 2ab\mathrm{Cov}(X,Y)$
Expectation and variance of the sum $Z_n = \sum_{i=1}^n a_i X_i$ of independent RV	$E(Z_n) = \sum_{i=1}^n a_i E(X_i)$ $V(Z_n) = \sum_{i=1}^n a_i^2 V(X_i)$
	$Z_n \approx N(\mu, \sigma^2)$ mit $\mu = E(Z_n),\ \ \sigma^2 = V(Z_n)$
Central Limit Theorem for the sum $Z_n = \sum_{i=1}^n a_i X_i$	Conditions: (i) all X_i independent from each other, (ii) variance exists, (iii) no variance is greater than $\sigma^2/30$. Otherwise, the X_i may be arbitrary (!!) discrete or continuous or mixed RV

Inferential statistics: estimators and test functions

Estimator for the expectation

$$\hat{\mu} = \bar{X}$$

Estimator for the true percentage

$$\hat{\vartheta}_i = f_i = \frac{h_i}{n}$$

Estimator for the variance

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-p}\sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ or } \frac{n}{n-p}\sum_{k=1}^{K} f_k(x_k^* - \bar{x})^2$$

p is the number of estimated parameters, often $p = 1$ (the expectation has been estimated by $\bar{X}$)

Estimator for the parameters of simple linear regression $Y = aX + b$

$$\begin{aligned} \hat{a} &= \bar{Y} - \hat{b}\bar{X} \\ \hat{b} &= \frac{\sum_{i=1}^{n}(X_i Y_i) - \bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - \bar{X}^2}, \end{aligned}$$

Test function for hypotheses on the expectation (boundary of H_0 at μ_0, known variance)

$$T = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim N(0;1)$$

(standard normal distribution)

Test of the true percentage ϑ_0 (boundary of H_0)

$$T = \frac{f - \vartheta_0}{\sqrt{\vartheta_0(1-\vartheta_0)}}\sqrt{n}\sqrt{\frac{N-1}{N-n}} \sim N(0;1)$$

(correction factor for finite populations $\sqrt{\frac{N-n}{N-1}}$ if the sample size n is not $\ll$ the population size N)

Test function for hypotheses on the expectation μ_0 (unknown variance)

$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}}\sqrt{n} \sim T(n-1)$$

(student-t distribution with $n-1$ degrees of freedom)

Test function for hypotheses on a regression parameter β_j (boundary of H_0 at β_{j0})

$$T = \frac{\bar{\hat{\beta}}_j - \beta_{j0}}{\sqrt{\hat{V}(\hat{\beta})}} \sim T(n-p)$$

(student-t distribution with $n-p$ degrees of freedom, p is the total number of estimated regression parameters, $\hat{V}(\hat{\beta})$ is the estimated variance of the LSE parameter estimator)

Test function for the variance test $\sigma^2 = \sigma_0^2$ (tests the ratio)

$$T = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2(n-1)$$

(chi-squared distribution with $n-1$ degrees of freedom)

Test of the correlation coefficient $\rho(x,y)$ for $\rho = 0$

$$T_\rho = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}}} \sim T(n-2)$$

Inferential statistics: Confidence intervals

Confidence intervall (CI) for known variance

$$\mu \in \bar{x} \pm z_{1-\alpha/2} \, \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$z_{1-\alpha/2}$	quantile of the standard normal distribution for $p = 1 - \alpha/2$,
α	error probability (e.g., $\alpha = 5\% \Rightarrow z_{1-\alpha/2} = z_{0.975} = 1.96$)

Confidence interval for the true percentage

$$\vartheta \in f \pm z_{1-\alpha/2} \, \sqrt{\frac{f(1-f)}{n}} \sqrt{\frac{N-n}{N-1}}$$

where f denotes the sampled relative frequency. Condition: $nf(1-f) \geq 9$

Confidence interval for unknown variance

$$\mu \in \bar{x} \pm t^{(n-1)}_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

where $t^{(n-1)}_p$ denotes the tabulated quantile of the student-t distribution with $(n-1)$ degrees of freedom

Confidence interval for the simple regression slope parameter b

$$b \in \hat{b} \pm t^{(n-2)}_{1-\alpha/2} \, \hat{\sigma}_b, \quad \hat{\sigma}_b^2 = \frac{\hat{\sigma}_R^2}{n s_x^2}, \quad \hat{\sigma}_R^2 = \frac{1}{n-2} \sum_{i=1}^{n} \Big(y_i - \hat{y}(x_i) \Big)^2$$

Confidence interval of the regression function $y = a + bx$ itself

$$y \in \hat{a} + \hat{b}x \pm t^{(n-2)}_{1-\alpha/2} \, \frac{\hat{\sigma}_R}{\sqrt{n}} \sqrt{1 + \frac{(x-\bar{x})^2}{s_x^2}}$$

Inferential statistics: parametric tests

One-sided interval test for ">" or "≥" for location parameters or correlations

H_0: $\mu > \mu_0$, $\vartheta > \vartheta_0$, $\beta_j > \beta_{j0}$, or $\rho_{xy} > 0$ can be rejected at error probability α if $t_{\text{data}} < -t_{1-\alpha}$ with $t_{1-\alpha}$ the $p = 1-\alpha/2$ quantile of the corresponding (standard normal or student-t) distribution (see "estimators and test functions")

One-sided interval test for "<" or "≤"

H_0: $\mu < \mu_0$, $\vartheta < \vartheta_0$, $\beta_j < \beta_{j0}$, or $\rho_{xy} < 0$ can be rejected at error probability α if $t_{\text{data}} > t_{1-\alpha}$ (test functions as above).

Two-sided point test

H_0: $\mu = \mu_0$, $\vartheta = \vartheta_0$, $\beta_j = \beta_{j0}$, or $\rho_{xy} = 0$ can be rejected at error probability α if $|t_{\text{data}}| > t_{1-\alpha/2}$ (test functions as above).

Tests for the variance

The test function $T = \hat{\sigma}^2/\sigma_0^2$ is chi-squared distributed under H_0. Since this function is not symmetric, also quantiles for $p < 0.5$ are tabulated, hence

- H_0: $\sigma > \sigma_0$ can be rejected if $t_{\text{data}} < t_\alpha$,
- H_0: $\sigma < \sigma_0$ can be rejected if $t_{\text{data}} > t_{1-\alpha}$,
- H_0: $\sigma = \sigma_0$ can be rejected if $t_{\text{data}} > t_{1-\alpha/2}$ OR $t_{\text{data}} < t_{\alpha/2}$

Inferential statistics: Non-parametric tests

Chi-squared goodness-of-fit test
H_0: data are consistent with distribution F_0

$$T = \sum_{k=1}^{K} \left(\frac{(h_k - h_k^e)^2}{h_k^e} \right) = \sum_{k=1}^{K} \left(\frac{h_k^2}{h_k^e} \right) - n \sim \chi^2(K-1-r)$$

r	#parameters to be estimated,
$n = \sum_k h_k$	#observations,
K	#classes,
h_k^e	expected absolute frequency h_k if H_0

Test decision

H_0 can be rejected if $t_{\text{data}} > t_{1-\alpha}^{(K-1-r)}$

Independence test
H_0: X and Y are independent

$$T = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} \left(\frac{(h_{ij} - h_{ij}^{\mathrm{e}})^2}{h_{ij}^{\mathrm{e}}} \right) = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} \left(\frac{h_{ij}^2}{h_{ij}^{\mathrm{e}}} \right) - n \sim \chi^2(m)$$

K_x, K_y	#classes,
$m = (K_x - 1)(K_y - 1)$	#degrees of freedom,
$h_{ij}^{\mathrm{e}} = \frac{h(x_i)h(y_j)}{n}$	expected absolute frequency if H_0
$h(x_i), h(y_j)$	sum over columns and rows
$n = \sum_j h(y_j)$	number of data points (X_i, Y_i)

Test for identical populations/distributions
H_0: Two or more samples are consistent with identical (but otherwise unspecified) population distributions

T as for the independence tests with

K_x	#classes,
$K_y = M$	#samples

Conditions for all nonparametric chi-squared tests

$h_k^e \geq 5$ for all classes k

Kolmogorow-Smirnow goodness-of fit test (KS test)

$$D = \max_x \left| F(x) - F^{(0)}(x) \right| \sim D(n)$$

$F(x)$	sample distribution function,
$F^{(0)}(x)$	distribution function if H_0
n	smaple size

KS test decision

H_0 can be rejected if $d_{\text{data}} > d_{n,1-\alpha} \approx \dfrac{c(\alpha)}{\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}}$

with

α	0.010	0.025	0.050	0.100
$c(\alpha)$	1.628	1.480	1.358	1.224

Standard normal distribution $\Phi(z)$
(symmetry: $\Phi(-z) = 1 - \Phi(z)$)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Quantiles $z_p = \Phi^{-1}(p)$ of the standard normal distribution $\Phi(z)$
(symmetry: $z_p = -z_{1-p}$)

$q = 0.60$	0.70	0.80	0.90	0.95	0.975	0.990	0.995	0.999	0.9995
0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Quantiles $t_{n,p}$ of the Student t distribution with n degrees of freedom (symmetry: $t_{n,p} = -t_{n,1-p}$; limit $t_{n,p} \to z_p$ for $n \to \infty$)

n	$p = 0.60$	0.70	0.80	0.90	0.95	0.975	0.990	0.995	0.999	0.9995
1	0.325	0.727	1.376	3.078	6.315	12.706	31.821	63.657	318.31	636.62
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	3.153	3.707	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.154	4.587
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Quantiles $\chi^2_{n,p}$ of the χ^2 distribution with n degrees of freedom (symmetry: $\chi^2_{n,p} = -\chi^2_{n,1-p}$; limit for large n: $\chi^2_{n,p} \approx n + \sqrt{2n}\, z_p$)

n	$p = 0.9900$	0.9750	0.9500	0.9000	0.8000	0.5000	0.2000	0.1000	0.05000
1	6.635	5.034	3.821	2.706	1.656	0.4589	0.06540	0.01638	0.004230
2	9.210	7.378	5.991	4.605	3.219	1.386	0.4463	0.2107	0.1026
3	11.34	9.348	7.815	6.251	4.642	2.366	1.005	0.5843	0.3518
4	13.28	11.15	9.488	7.779	5.989	3.357	1.649	1.064	0.7106
5	15.09	12.83	11.07	9.236	7.289	4.351	2.343	1.610	1.155
6	16.81	15.45	12.59	10.64	8.558	5.348	3.070	2.204	1.635
7	18.48	16.01	15.07	12.02	9.803	6.346	3.822	2.833	2.167
8	20.10	17.54	15.51	13.36	11.03	7.344	4.594	3.490	2.733
9	21.67	19.03	16.92	15.68	12.24	8.343	5.380	4.168	3.325
10	23.22	20.49	18.31	15.99	13.44	9.342	6.179	4.865	3.940
15	30.59	27.49	25.0	22.31	19.31	15.34	10.31	8.547	7.261
20	37.58	34.18	31.41	28.41	25.04	19.34	15.58	12.44	10.85
30	50.92	46.99	43.78	40.26	36.25	29.34	23.36	20.60	18.49