$\sum_i n_i = n.$ absolute Haäufigkeit n_i

$f_i = \frac{n_i}{n}$ relativen Häufigkeiten

$x_{(1)} \le x_{(2)} \le \dots x_{(n-1)} \le x_{(n)}$

range $= x_{(n)} - x_{(1)}$ (Spannweite)

Lower bound $= Q_1 - 1.5 \cdot \mathrm{IQR}$

Upper bound $= Q_3 + 1.5 \cdot \mathrm{IQR}$

Interquartile range $\mathrm{IQR} = Q_3 - Q_1$

$$\text{Median} = \begin{cases} x_{(k+1)}, & n = 2k+1 \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}), & n = 2k, \end{cases}$$

Sample mean (Mittelwert)

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Geometric mean (geometrische Mittel)

$$\bar{x}_n^{(g)} = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$$

Harmonic mean (harmonische Mittel)

$$\bar{x}_n^{(h)} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Laplace space For $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, $|\Omega| = n$

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{number of elements of } A}{\text{number of elements of } \Omega} \qquad p_i = P(\{\omega_i\}) = \frac{1}{|\Omega|} = \frac{1}{n}$$

Geometric probabilities

$$P(A) = \frac{\text{area of desired outcomes}}{\text{area of total outcomes}}$$

Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplication theorem

- For $P(B) > 0$ it holds $P(A \cap B) = P(A|B) \cdot P(B)$
- For $P(A) > 0$ it holds $P(A \cap B) = P(B|A) \cdot P(A)$
- For $A_1, A_2, \dots, A_n \in \mathcal{A}$ with $P(A_1 A_2 \dots A_n) > 0$ holds $P(\bigcap_{i=1}^{n} A_i) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1A_2) \dots P(A_n|A_1A_2 \dots A_{n-1})$

Independence $P(A|B) = P(A)$ $\quad P(A \cap B) = P(A) \cdot P(B)$

we say that the events A and B are independent.

Bayes' law

$$P(C_i|A) = \frac{P(A|C_i) \cdot P(C_i)}{\sum_{j=1}^{\infty} P(A|C_j) \cdot P(C_j)}$$

Bayes' law based on the odds

$$\underbrace{\frac{P(H|A)}{P(H^c|A)}}_{\text{a posteriori odds}} = \underbrace{\frac{P(H)}{P(H^c)}}_{\text{a priori odds}} \cdot \underbrace{\frac{P(A|H)}{P(A|H^c)}}_{\text{likelihood quotient}}$$

Empirical Rule

	samples	populations
Approximately 68%	$(\bar{x} - s, \bar{x} + s)$	$(\mu - \sigma, \mu + \sigma)$
Approximately 95%	$(\bar{x} - 2s, \bar{x} + 2s)$	$(\mu - 2\sigma, \mu + 2\sigma)$
Approximately 99.7%	$(\bar{x} - 3s, \bar{x} + 3s)$	$(\mu - 3\sigma, \mu + 3\sigma)$

Independence

$$E(XY) = \int_a^b \int_c^d x \cdot y \cdot f(x,y)\, dx dy$$

$$E(XY) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i \cdot y_j \cdot p(x_i, y_j).$$

Two random variables X and Y are independent if

$$E(XY) = E(X) \cdot E(Y).$$

Covariance

$$Cov(X,Y) = E\left((X - E(X)) \cdot (Y - E(Y))\right).$$

- $Cov(aX + b, cY + d) = ac\, Cov(X,Y)$ for constants a, b, c and d.
- $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$
- $Cov(X, X) = Var(X)$
- $Cov(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$ for any X and Y
- If X and Y are independent, then $Cov(X,Y) = 0$.

Correlation $\quad -1 \le \rho(X,Y) \le 1$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}.$$

Empirical central moment of order r

$$m_r = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}_n)^r$$

Sample variance (Stichprobenstreuung)

$$s_n^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}_n^2\right)$$

Correlation coefficient $-1 \le r_{xy} \le 1$

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right) \cdot \left(\frac{y_i - \bar{y}}{s_y}\right)$$

empirical covariance is $s_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$.

Sample space (Merkmalraum) Ω

Event (Ereignis) $A \subseteq \Omega$ $\quad P(A) = \sum_{\omega \in A} P(\omega)$

Measurable space (Messraum) $(\Omega, \mathcal{A})$

Probability space (Wahrscheinlichkeitsraum) $(\Omega, \mathcal{A}, P)$

rules of probability

(1) $P(A^c) = 1 - P(A)$

(2) If B and C are disjoint then $P(B \cup C) = P(B) + P(C)$

(3) If B and C are not disjoint, we have the *inclusion-exclusion principle*

$$P(B \cup C) = P(B) + P(C) - P(B \cap C)$$

Probability mass function $\quad p(a) = P(X = a).$

Cumulative distribution function $\quad F(a) = P(X \le a).$

Bernoulli distribution $\quad Bernulli(p)$

X takes the values 0 and 1

$P(X = 1) = p$ and $P(X = 0) = 1 - p$

Binomial distribution $\quad B(n,p)$

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

Geometric distribution

$$p(k) = P(X = k) = (1-p)^k \cdot p.$$

Poisson distribution $\quad P(\lambda)$

$$p(k) = P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

Hypergeometric

$$p(k) = P(X = k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}} \text{ for } k = \max\{0, n-(N-r)\}, \dots, \min\{r, n\}.$$

Exponential distribution $\quad \exp(\lambda)$

$$f(x) = \lambda\, e^{-\lambda x}, \quad \text{for } x \ge 0.$$

$$F(x) = P(X \le x) = \int_0^x \lambda e^{-\lambda t}\, dt = -e^{-\lambda t}\Big|_0^x = 1 - e^{-\lambda x}, \quad x \ge 0$$

$$E(X) = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2} \qquad \sigma = \sqrt{Var(X)} = \frac{1}{\lambda}$$

Uniform distribution $\quad \mathcal{U}(a,b)$

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a,b) \\ 0, & \text{otherwise} \end{cases}$$

Gamma distribution $\quad Gamma(\alpha, \beta)$

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\, \beta^\alpha}, \quad \text{for } x > 0.$$

Gamma function $\quad E(X) = \alpha\beta \quad Var(X) = \alpha\beta^2$

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u}\, du \quad \text{for } \alpha > 0$$

$$F(x) = P(X \le x) = \int_0^x \frac{t^{\alpha-1} \cdot e^{-\frac{t}{\beta}}}{\Gamma(\alpha)\, \beta^\alpha}\, dt, \quad x \ge 0$$

$\chi^2(n)$ $\quad P(X \le \chi^2_{n;p}) = p, \quad$ for $p \in (0,1)$ $\quad \chi^2(n)$

$$f(x) = \frac{x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}}{\Gamma(\frac{n}{2}) \cdot 2^{\frac{n}{2}}} \quad \text{for } x > 0.$$

Normal (Gaussian) distribution $\quad \mathcal{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \Phi(x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}\, dt$$

Skewness

$$g_1^{(1)} = \frac{m_3}{m_2^{\frac{3}{2}}} = \frac{\sqrt{n}}{(n-1)^{\frac{3}{2}}} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3$$

$$g_1^{(2)} = \frac{n}{(n-1)(n-2)} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3$$

$$g_1^{(3)} = \frac{m_3}{s^3} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3$$

$g_1 > 0$ negative skewed : *mode* $<$ *median* $< \bar{x}$

$g_1 < 0$ positive skewed : *mode* $>$ *median* $> \bar{x}$

Kurtosis

$$g_2^{(1)} = \frac{m_4}{m_2^2} = \frac{\sqrt{n}}{(n-1)^2} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^4$$

$$g_2^{(2)} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^4$$

$$g_2^{(3)} = \frac{m_4}{s^4} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^4$$

The expected value or mean of X is defined by

$$E(X) = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + \dots x_n \cdot p(x_n) = \sum_{i=1}^{n} x_i \cdot p(x_i).$$

$$E(X + Y) = E(X) + E(Y)$$

$$E(aX + b) = aE(X) + b$$

$$E(h(X)) = \sum_{i=1}^{n} h(x_i)\, p(x_i).$$

Variance and standard deviation

$$Var(X) = E((X - \mu)^2) \qquad \sigma = \sqrt{Var(X)}.$$

$$= E(X^2) - \mu^2$$

$$Var(X) = \sum_{k=1}^{n} p(x_k) \cdot (x - \mu)^2$$

$$Var(X + b) = Var(X)$$

$$Var(aX) = a^2 Var(X)$$

$Var(X) = 0$ then X is constant.

X and Y be independent

$$Var(X + Y) = Var(X) + Var(Y)$$

Expected value and variance

$$\mu = E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx$$

$$Var(X) = E((X - E(X))^2) = \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x)\, dx$$

$$\sigma = \sqrt{Var(X)}$$

Probability density function

$$P(c \le x \le d) = \int_c^d f(x)\, dx$$

Cumulative distribution function

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\, dt$$

Standardization $Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$

normal statistic $\quad Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$

Kurtosis

$$k = \frac{E(X - E(X}{\sigma^4}$$

- $k = 3$ - Normal
- $k > 3$ - more "peaked"
- $k < 3$ more "flat"

Confidence intervals for μ:

Normal z- statistic

$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$ $\quad Z = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0,1)$

Let $X_1, \dots, X_n$ be a large sample from $X \sim \mathcal{N}(\mu, \sigma^2)$, with known σ

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

$$P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2} \qquad z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \tfrac{\alpha}{2})$$

Let $X_1, \dots, X_n$ be a large sample from $\mathcal{N}(\mu, \sigma^2)$, with unknown σ

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]$$

$$P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2} \qquad z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \tfrac{\alpha}{2})$$

Student's t- statistic

$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1},$

Let $X_1, \dots, X_n$ be a small sample from $\mathcal{N}(\mu, \sigma^2)$, with unknown σ

$$\left[\bar{x} - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}\right]$$

$$P(t > t_{\frac{\alpha}{2},n-1}) = \frac{\alpha}{2}$$

χ^2-distribution

$\chi^2 = \frac{s^2}{\frac{\sigma^2}{n-1}} = \frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2_{n-1}$

Let the data $x_1, \dots, x_n$ is drown from $\mathcal{N}(\mu, \sigma^2)$ with mean μ and standard deviation σ both unknown.

$$\frac{(n-1)\cdot s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)\cdot s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$$

$$P(\chi^2 > \chi^2_{\alpha,n}) = \alpha$$

Small independent samples for $\mu_1 - \mu_2$: t-statistic

$$\left[(\bar{x}_1 - \bar{x}_2) - t^* \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \bar{x}_1 - \bar{x}_2) + t^* \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

$$s_p = \sqrt{\frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 + n_2 - 2}} \qquad t^* = t_{n_1+n_2-2,\frac{\alpha}{2}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

H_1	Rejection region	
$\mu_1 < \mu_2$	$t < -t_{n_1+n_2-2,\alpha}$	$P(t > t_{n_1+n_2-2,\alpha}) = \alpha$
$\mu_1 > \mu_2$	$t > t_{n_1+n_2-2,\alpha}$	$P(t > t_{n_1+n_2-2,\alpha}) = \alpha$
$\mu_1 \neq \mu_2$	$\lvert t\rvert > t_{n_1+n_2-2,\frac{\alpha}{2}}$	$P(\lvert t\rvert > t_{n_1+n_2-2,\frac{\alpha}{2}}) = \alpha$

Sampling distribution of $\hat{p}_1 - \hat{p}_2$

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 \quad \text{and} \quad \sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(p_1 - p_2,\ \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

Confidence interval for $p_1 - p_2$

$$\left[(\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},\ (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right]$$

Hypothesis testing about $p_1 - p_2$

$$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}}$$

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1}{n_1 + n_2}\hat{p}_1 + \frac{n_2}{n_1 + n_2}\hat{p}_2.$$

The F distribution

critical value

$$F_{\frac{\alpha}{2},df1,df2} = F_{0.05,12,17}$$

$$F = \frac{\text{larger sample variance}}{\text{smaller sample variance}} = \frac{s_1^2}{s_2^2}$$

One-way ANOVA — Two estimates of σ^2

$$F = \frac{\text{variance between samples}}{\text{variance within samples}} = \frac{MSB}{MSW}$$

the total sum of squares

$SST = SSB + SSW =$

the between-sample sum of squares SSB

$$SSB = \sum_i \frac{T_i^2}{n_i} - \frac{(\Sigma x)^2}{n}$$

the within-samples sum of squares SSW

$$SSW = \Sigma x^2 - \sum \frac{T_i^2}{n_i}$$

The variance between samples MSB and the variance within samples MSW are calculated as

$$\text{MSB} = \frac{SSB}{k-1} \qquad \text{MSW} = \frac{SSW}{n-k}$$

where $k-1$ and $n-k$ are the df for the numerator and the df for the denominator for the F-distribution, and k is the number of samples.

$x =$ the score of a student

$k =$ the number of different samples

$n_i =$ the size of sample i

$T_i =$ the sum of the values in sample i

$n =$ the number of values in all samples; $n = \Sigma_i n_i$

$\Sigma x =$ the sum of the values in all samples; $\Sigma x = \Sigma_i T_i$

$\Sigma x^2 =$ the sum of the squares of the values in all samples

sampling distribution of $\hat{p}$

$$\hat{p} = \frac{X}{n} \qquad X \sim B(n, p)$$

$$E(\hat{p}) = p \quad \text{and} \quad Var(\hat{p}) = \frac{p\,q}{n}$$

For large samples, the sampling distribution is approximately normal

$$\hat{p} \approx \mathcal{N}\left(p, \frac{pq}{n}\right)$$

$$\left[\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\,\hat{q}}{n}},\ \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\,\hat{q}}{n}}\right] \qquad \hat{p} = \frac{x}{n} \text{ and } \hat{q} = 1 - \hat{p}$$

Adjusted confidence interval for p

$$\left[\tilde{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\tilde{p}\,\tilde{q}}{n+4}},\ \tilde{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\tilde{p}\,\tilde{q}}{n+4}}\right] \qquad \tilde{p} = \frac{x+2}{n+4} \qquad \tilde{q} = 1 - \tilde{p}.$$

Polling confidence interval

Bernulli(p) distribution, where p is unknown

Conservative normal

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}},\ \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}}\right]$$

Rule-of-thumb 95% confidence interval for p

$$\left[\bar{x} - \frac{1}{\sqrt{n}},\ \bar{x} + \frac{1}{\sqrt{n}}\right]$$

Choosing a sample size

$$ME = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\,\hat{q}}{n}} \le m \qquad \text{margin of error}$$

Sampling distribution of $\bar{x}_1 - \bar{x}_2$

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

$$\bar{x}_1 - \bar{x}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Confidence interval for $\mu_1 - \mu_2$

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

Pooled sample estimator of σ^2

$$\sigma_1^2 = \sigma_2^2 = \sigma^2 \qquad s_p^2 = \frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 + n_2 - 2}$$

Pearson's Chi-square statistic

goodness of fit

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

degrees of freedom is $df = k - 1$

$O =$ observed frequency for a category

$E =$ expected frequency for a category $= np$

$O \quad p \quad E = np \quad (O - E)^2 \quad \frac{(O-E)^2}{E}$

Test of independence

$df = (R - 1) \cdot (C - 1)$

$$E = \frac{(\text{Row total}) \cdot (\text{Column total})}{\text{Sample size}}$$

ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between	$k - 1$	SSB	MSB	$F = \frac{\text{MSB}}{\text{MSW}}$
Within	$n - k$	SSW	MSW	
Total	$n - 1$	SST		

Regression and Correlation

$$r^2 = \frac{SSR}{SST} = \frac{\text{model's variation}}{\text{total variation}} = \text{"coefficient of deermination"}.$$

Analysis of variance (ANOVA)

$$F = \frac{MSR}{MSE} = \frac{SSR}{s^2}$$

Source of variation	SS	df	MS	F	p
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	p-value
Error	SST	$n - 2$	$MSE = \frac{SSE}{n-2}$		
Total	SST	$n - 1$			

Test statistic

(t-statistic analog)

$H_0 : \mu = \mu_0$

$H_1 : \mu < \mu_0$

$H_1 : \mu > \mu_0$

$H_1 : \mu \neq \mu_0$

If the p-value $< \alpha$ we reject H_0

If the p-value $> \alpha$ we do not reject H_0

Truth	Decision: fail to reject H_0	Decision: reject H_0
H_0 true	✓	Type 1 Error
H_1 true	Type 2 Error	✓

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad p\text{-value} = P(Z < z)$$

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad p\text{-value} = P(Z > z)$$

$$p\text{-value} = P(|Z| > z)$$

	H_1	Rejection region	
$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$\mu < \mu_0$	$(-\infty,\ -z_\alpha)$	$P(Z > z_\alpha) = \alpha$
	$\mu > \mu_0$	$(z_\alpha,\ +\infty)$	$P(Z > z_\alpha) = \alpha$
	$\mu \neq \mu_0$	$(-\infty,\ z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, +\infty)$	$P(\lvert Z\rvert > z_{\frac{\alpha}{2}}) = \alpha$
$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$	$\mu < \mu_0$	$(-\infty,\ -t_{n-1,\alpha})$	$P(t > t_{n-1,\alpha}) = \alpha$
	$\mu > \mu_0$	$(t_{n-1,\alpha},\ +\infty)$	$P(t > t_{n-1,\alpha}) = \alpha$
	$\mu \neq \mu_0$	$(-\infty,\ t_{n-1,\frac{\alpha}{2}}) \cup (t_{\frac{\alpha}{2}}, +\infty)$	$P(t > t_{n-1,\frac{\alpha}{2}}) = \alpha$

Errors, significance level and power

Significance level $= P(\text{Type 1 Error})$
$= \text{probability we incorrectly reject } H_0$
$= P(\text{test statistic in rejection region} \mid H_0)$
$= \alpha$

Power $= \text{probability we correctly reject } H_0$
$= P(\text{test statistic in rejection region} \mid H_1)$
$= 1 - P(\text{Type 2 Error})$
$= 1 - \beta$

Hypothesis testing for a population proportion

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Test of hypothesis about a population variance

$$\chi^2 = \frac{(n-1)\,s^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

H_1	Rejection region	
$\sigma^2 < \sigma_0^2$	$(0,\ \chi^2_{n-1,1-\alpha})$	$P(\chi^2 > \chi^2_{n-1,\alpha}) = 1 - \alpha$
$\sigma^2 > \sigma_0^2$	$(\chi^2_{n-1,\alpha},\ +\infty)$	$P(\chi^2 > \chi^2_{n-1,\alpha}) = \alpha$
$\sigma^2 \neq \sigma_0^2$	$(0,\ \chi^2_{n-1,1-\frac{\alpha}{2}}) \cup (\chi^2_{n-1,\frac{\alpha}{2}}, +\infty)$	$P(\chi^2 > \chi^2_{n-1,\frac{\alpha}{2}}) = \frac{\alpha}{2}$

Hypothesis testing about $\mu_1 - \mu_2$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \qquad z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

H_1	Rejection region	
$\mu_1 < \mu_2$	$(-\infty,\ -z_\alpha)$	$P(Z > z_\alpha) = \alpha$
$\mu_1 > \mu_2$	$(z_\alpha,\ +\infty)$	$P(Z > z_\alpha) = \alpha$
$\mu_1 \neq \mu_2$	$(-\infty,\ z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, +\infty)$	$P(\lvert Z\rvert > z_{\frac{\alpha}{2}}) = \alpha$