# Formulae Collection – Modul 42073

## Descriptive measures

Arithmetic mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$p^{\text{th}}$-quantile: $x_p = \begin{cases} x_{(\lfloor np \rfloor + 1)} & \text{for } np \notin \mathbb{N} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}) & \text{for } np \in \mathbb{N} \end{cases}$

Median: $x_{med} = x_{0.5}$

Pearson's corrected contingency coefficient: $K_P^* = \left(\frac{\chi^2}{\chi^2+n}\right)^{\frac{1}{2}} \cdot \left(\frac{\min(k,l)}{\min(k,l)-1}\right)^{\frac{1}{2}}$

Sample correlation: $r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$

## Selecting $k$ from $n$ objects

| | without replacement ($k \le n$) | with replacement |
|---|---|---|
| withour order | $\binom{n}{k}$ | $\binom{n+k-1}{k}$ |
| with order | $\frac{n!}{(n-k)!}$ | $n^k$ |

## Rules of probability

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

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

Bayes' Rule: $P(B_k|A) = \frac{P(B_k) \cdot P(A|B_k)}{\sum\limits_{i=1}^{m} P(B_i)P(A|B_i)}$

## Random variables and distributiosn

| | discrete | continuous |
|---|---|---|
| density | $f(x) = P(X = x)$ | $f(x) \ge 0,\ \int_{-\infty}^{\infty} f(x)dx = 1$ |
| CDF | $F(x) = P(X \le x) = \sum_{t \le x, f(t)>0} f(t)$ | $F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$ |
| E(X) | $E(X) = \sum_{x \in \Omega} x f(x)$ | $E(X) = \int_{-\infty}^{\infty} x f(x)dx$ |
| $\mathsf{Var}(X)$ | $\mathsf{Var}(X) = \sum_{x \in \Omega}(x - E(X))^2 f(x)$ | $\mathsf{Var}(X) = \int_{-\infty}^{\infty}(x - E(X))^2 f(x)dx$ |

| **Name** | **Notation** | **f(x)** | **Support $\Omega$** | **E(X)** | **Var(X)** |
|---|---|---|---|---|---|
| **Discrete distributions** | | | | | |
| Binomial | $X \sim Bin(n,p)$ | $\binom{n}{x}p^x(1-p)^{n-x}$ | $x \in \{0,1,\ldots,n\}$ | $np$ | $np(1-p)$ |
| Hypergeometric | $X \sim \mathcal{H}(N,M,n)$ | $\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$ | $0 \le x \le M$, $0 \le n-x \le N-M$ | $\frac{nM}{N}$ | $n\frac{M}{N}(1-\frac{M}{N})\frac{N-n}{N-1}$. |
| Discrete uniform | $X \sim DU(m)$ | $\frac{1}{m}$ | $\{1,\ldots,m\}$ | $\frac{m+1}{2}$ | $\frac{m^2-1}{12}$ |
| Poisson | $X \sim Poi(\lambda)$ | $\exp(-\lambda)\frac{\lambda^x}{x!}$ | $x \in \mathbb{N}_0$ | $\lambda$ | $\lambda$ |
| **Continuous distributions** | | | | | |
| Normal | $X \sim N(\mu,\sigma^2)$ | $\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ | $x \in \mathbb{R}$ | $\mu$ | $\sigma^2$ |
| Standard normal | $X \sim N(0,1)$ | $\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$ | $x \in \mathbb{R}$ | $0$ | $1$ |
| Continuous uniform | $X \sim Unif(a,b)$ | $\frac{1}{b-a}$ | $x \in [a,b]$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| Exponential | $X \sim Exp(\lambda)$ | $\lambda\exp(-\lambda x)$ | $x \ge 0$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |

## Confidence intervals

$100(1-\alpha)\%$-**confidence interval for mean** $E(X)$

| | |
|---|---|
| $X \sim N(\mu,\sigma^2)$ or $n$ large ($n \ge 30$), $\sigma$ known | $\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\quad \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$ |
| $X \sim N(\mu,\sigma^2)$, $\sigma$ unknown | $\left[\bar{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}},\quad \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$ |
| $n$ large ($n \ge 30$), $\sigma$ unknown | $\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}},\quad \bar{X} + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$ |

$100(1-\alpha)\%$-**confidence interval for proportion** $p$

| | |
|---|---|
| $X \sim Bin(1,p)$, n large ($n \ge 30$) | $\left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\quad \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$ with $\hat{p} = \bar{X}$ |

Prof. Dr. Sonja Kuhnt, Prof. Dr. Nadja Bauer

## Hypothesis testing

| Null hypothesis | Alternative hypothesis | Test statistic | Critical region |
|---|---|---|---|
| **(Approximate) z-test on the mean** ($X \sim N(\mu, \sigma^2)$ or $n \geq 30$, $\sigma$ known) | | | |
| $\mu = \mu_0$ | $\mu \neq \mu_0$ | | $\lvert z\rvert > z_{1-\frac{\alpha}{2}}$ |
| $\mu \geq \mu_0$ | $\mu < \mu_0$ | $Z = \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ | $z < -z_{1-\alpha}$ |
| $\mu \leq \mu_0$ | $\mu > \mu_0$ | | $z > z_{1-\alpha}$ |
| **One sample t-test on the mean** ($X \sim N(\mu, \sigma^2)$, $\sigma$ unknown) | | | |
| $\mu = \mu_0$ | $\mu \neq \mu_0$ | | $\lvert t\rvert > t_{n-1,1-\frac{\alpha}{2}}$ |
| $\mu \geq \mu_0$ | $\mu < \mu_0$ | $T = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}$ | $t < -t_{n-1,1-\alpha}$ |
| $\mu \leq \mu_0$ | $\mu > \mu_0$ | | $t > t_{n-1,1-\alpha}$ |
| **Approximate z-test on the mean** ($n \geq 30$, $\sigma$ unknown, $E(X) = \mu$) | | | |
| $\mu = \mu_0$ | $\mu \neq \mu_0$ | | $\lvert z\rvert > z_{1-\frac{\alpha}{2}}$ |
| $\mu \geq \mu_0$ | $\mu < \mu_0$ | $Z = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}$ | $z < -z_{1-\alpha}$ |
| $\mu \leq \mu_0$ | $\mu > \mu_0$ | | $z > z_{1-\alpha}$ |
| **Two-sample t-test on a difference in mean** ($X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$, $\sigma_X$, $\sigma_Y$ unknown) | | | |
| $\mu_X - \mu_Y = \delta_0$ | $\mu_X - \mu_Y \neq \delta_0$ | | $\lvert t\rvert > t_{k,1-\frac{\alpha}{2}}$ |
| $\mu_X - \mu_Y \geq \delta_0$ | $\mu_X - \mu_Y < \delta_0$ | $T = \frac{\overline{X}-\overline{Y}-\delta_0}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}}$ | $t < -t_{k,1-\alpha}$ |
| $\mu_X - \mu_Y \leq \delta_0$ | $\mu_x - \mu_Y > \delta_0$ | | $t > t_{k,1-\alpha}$ |
| | | with $k = \left[\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2 \Big/ \left(\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2\right)\right]$ | |
| **Large sample test on a proportion** ($X \sim Bin(1,p)$, $n \geq 30$ ) | | | |
| $p = p_0$ | $p \neq p_0$ | | $\lvert z\rvert > z_{1-\frac{\alpha}{2}}$ |
| $p \geq p_0$ | $p < p_0$ | $Z = \frac{\bar{X}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ | $z < -z_{1-\alpha}$ |
| $p \leq p_0$ | $p > p_0$ | | $z > z_{1-\alpha}$ |
| **Chi-square independence test** ($X$ and $Y$ categorical, $\tilde{h}_{ij} > 5$) | | | |
| Variables $X$ and $Y$ are stochastically independent | Variables $X$ and $Y$ are stochastically dependent | $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(h_{ij}-\tilde{h}_{ij})^2}{\tilde{h}_{ij}}$, $\tilde{h}_{ij} = \frac{h_{i\bullet}\cdot h_{\bullet j}}{n}$ | $\chi^2 > \chi^2_{1-\alpha,(k-1)(l-1)}$ |

## Regression

Regression model: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $\quad$ $\epsilon_i \sim N(0, \sigma^2) \quad i = 1, \ldots, n$

Estimator: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$ $\quad$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2} \qquad \hat{\sigma}_{\widehat{\beta}_1} = \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \hat{\sigma}_{\widehat{\beta}_0} = \sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

Fitted regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ $\quad$ $(\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i)$

Coefficient of determination: $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ $\quad$ Residuals: $e_i = y_i - \hat{y}_i$

| Null hypothesis | Alternative hypothesis | Test statistic | Critical region |
|---|---|---|---|
| $\beta_1 = \beta_{1,0}$ | $\beta_1 \neq \beta_{1,0}$ | | $\lvert t\rvert > t_{n-2,1-\alpha/2}$ |
| $\beta_1 \geq \beta_{1,0}$ | $\beta_1 < \beta_{1,0}$ | $T = \frac{\widehat{\beta}_1-\beta_{1,0}}{\hat{\sigma}_{\widehat{\beta}_1}}$ | $t < -t_{n-2,1-\alpha}$ |
| $\beta_1 \leq \beta_{1,0}$ | $\beta_1 > \beta_{1,0}$ | | $t > t_{n-2,1-\alpha}$ |

**$100(1-\alpha)\%$-prediction interval on a future observation $Y_0$ at value $x_0$**

$$\left[\widehat{y}_0 - t_{n-2,1-\alpha/2}\sqrt{\widehat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}\,,\ \widehat{y}_0 + t_{n-2,1-\alpha/2}\sqrt{\widehat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}\right]$$