Chapter 3

Inferential Statistics

Inferential Statistics

- Informal aim: draw inferences about the population on the basis of the sample information.
- The conclusions/inverences can be incorrect. To minimize the probability of incorrect decisions we use the methods of probability theory.
- Formal aim: statements about the characteristics of the attribute X (random variable or random vector)
- All relevant information about X can be recovered from the distribution function F of X, i.e. $F(x) = P(X \leq x)$, since all necessary probabilities can be computed using $F(x)$.

Example:

- Delivery of 1000 pieces of a particular product; M pieces are defective
- M is unknown
- Random selection of $n = 30$ pieces („sample“)

  The sample contains 2 defective

  Possible aims:

  - Estimate M (e.g. $\frac{2}{30} \cdot 1000 = 66.67$)
  - Estimate an interval for M (e.g. $M \in [58; 84]$ with prob. of 95%)
  - Test hypothesis that $M > 50$.

Step 1: fixing the family of distributions

Problem: The distribution function F of the variable of interest X is unknown in general.

- We have some preliminary information about F and use it to choose the family of potential distributions.
- The family of distributions $\mathcal{F}$ is indexed by the parameter ϑ, i.e. $X \sim F_\vartheta$, $\vartheta \in \Theta$, where Θ is the set of parameter values.
- Thus $\mathcal{F} = \{F_\vartheta : \vartheta \in \Theta\}$.

Examples:

- Let X be the quality of produced bulbs (1 functioning, 0 defective).
  - Then $X \sim B(1, p)$ with $p = P(X = 1)$.
  - It holds $\vartheta = p$, $\Theta = (0, 1)$ and $\mathcal{F} = \{B(1, p) : p \in (0, 1)\}$.
- Let X be the body height. Many studies have shown that the body height is approximately normally distributed.
  - It holds $X \sim \mathcal{N}(\mu, \sigma^2)$.
  - Thus $\vartheta = (\mu, \sigma)$ and $\Theta = \mathbb{R} \times (0, \infty)$.

If $\Theta \subset \mathbb{R}^k$, then we speak about a parametric family of distributions.

Note:

- If X is a discrete RV, then the true distribution function is usually contained in the family (for example, the Bernoulli distribution)
- If X is a continuous random variable, then the true distribution function is, in general, not contained in the family.
  - The family is used as an approximation.
  - The choice of the family follows from the analysis of the histogram.
  - The choice should be statistically justified. Exact procedures: goodness-of-fit tests (later).

Example: The asset returns are frequently assumed to follow the normal distribution, i.e. $\mathcal{F} = \{N(\mu, \sigma^2)\}$. But if the returns are in fact t-distributed, then the family does not contain the correct distribution.

Step 2: draw a sample

- To make statements about the parameter ϑ or the distribution function F, we run random experiments.
- We draw from the population samples of size n. This leads to the sample $x_1, ..., x_n$.
- Method: $x_1, ..., x_n$ are seen as realizations of random variables $X_1, ..., X_n$ which are called sample variables.
- It is usually assumed that $X_1, ..., X_n$ are independent and follow the same distribution as X (identically distributed).

Example:

- Pick up a sample of delivered products and control the quality. The sample is, e.g. $1, 0, 0, 0, 1, 0$......
- Pick up a sample students and measure the height. The sample is, e.g. $1.65, 1.86, 1.73, 1.91, ...$

Step 3: Disciplines of the inferential statistics

a) Parameter estimation (point estimation)

Aim: Using the sample data we determine an estimator (value) for the unknown parameter ϑ

Example:

- estimate the fraction of the voters, who voted for a particular party, using an exit-poll
- expected lifetime of a particular product by running a selective quality control

b) Confidence intervals (interval estimation)

We determine the interval, where the true (but unknown) parameter lies with high probability.

Example: an interval, which contains the true fraction of voters who voted for a particular party and where this true fraction lies with probability of 95%.

c) Testing hypotheses

Would the party A reach (at least, at most) 20 % of votes, i. e. does it hold that $p \geq 0.20$ or $p \leq 0.20$?

Parameter estimation

Assumptions: Let $X \sim F_\vartheta$, $x_1, ..., x_n$ is a sample and $X_1, ..., X_n$ are sample variables.

- By test statistics we denote a (measurable) function, which depends only on the sample variables $X_1, ..., X_n$.
- A test statistics $g(X_1, \ldots, X_n)$, which is used to estimate the parameter ϑ, is called the estimation function or shortly the estimator of ϑ.
- $g(x_1, \ldots, x_n)$ is the estimate of ϑ.

Notation: we denote the estimator/estimate of ϑ by $\hat{\vartheta}$.

Examples:

- Expectation $E(X)$
  - $E(X) = \sum_i x_i P(X = x_i)$ or $E(X) = \int_{-\infty}^{\infty} x f(x) dx$
- Since a-priori each value has the same probability, we obtain the estimator by replacing $P(X = i)$ with $1/n$.

Estimator for $E(X)$

Estimator: $\hat{\vartheta} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Estimate: sample mean $\hat{\vartheta} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

- Variance
  - $Var(X) = \sum_i (x_i - E(X))^2 P(X = x_i)$ or
    $Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$

Estimator for $Var(X)$

Estimator: $\hat{\vartheta} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$

Estimate: empirical variance $\hat{\vartheta} = \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$

Universität Augsburg

Examples:

- Distribution function F: the empirical distribution function is an estimate/or for F.
- Density or probability function f: the histogram is an estimate/or of f.

We have a link between the characteristics of a data set (compare descriptive statistics) and the characteristics of a distribution function. The characteristics of a data set are estimators of the corresponding characteristics of the distribution function!!!

Maximum Likelihood (ML) estimation

Given:

- A random sample $(x_1, \ldots, x_n)$
- Likelihood function $L_\vartheta(x_1, \ldots, x_n)$: the joint density function $f(x_1, \ldots, x_n|\vartheta)$

$$f(x_1, \ldots, x_n|\vartheta) = \prod_{i=1}^{n} f_\vartheta(x_i) = L_\vartheta(\underline{x})$$

If X is a discrete RV, then $L_\vartheta(x_1, \ldots, x_n)$ is the probability that we observe the particular sample assuming that the underlying parameter equals ϑ.

Typical ML-estimation

1. Write down the likelihood function : $L_\vartheta(x_1, \ldots, x_n)$
2. Take logs (optional): $\ln L_\vartheta(x_1, \ldots, x_n)$
3. Take derivatives: $\frac{\partial}{\partial\vartheta}[\ln]L_\vartheta(x_1, \ldots, x_n) \stackrel{!}{=} 0$
4. (usually thete are many parameters!)

Example: Bernoulli distribution

Success of a therapy: $X = \begin{cases} 1 & \text{if successful} \\ 0 & \text{if not} \end{cases} \sim B(1,p), p \in (0,1)$

The sample variables $X_1, ..., X_n$ are independent and $X_i \sim B(1,p)$ for $i = 1, .., n$.

Since $P_p(X_i = x_i) = f_i(x_i) = p^{x_i}(1-p)^{1-x_i}$ für $x_i \in \{0,1\}$ it holds that

$$L_p(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i) = \prod_{i=1}^{n} p^{x_i}\,(1-p)^{1-x_i}$$

$$= p^{\sum\limits_{i=1}^{n} x_i}\,(1-p)^{\sum\limits_{i=1}^{n}(1-x_i)} = p^{n\,\bar{x}}\,(1-p)^{n-n\,\bar{x}}$$

Taking logarithm yields

$$\ln L_p(x_1, \ldots, x_n) = n\,\bar{x}\,\ln p + (n - n\,\bar{x})\,\ln(1-p)$$

Universität Augsburg

Maximum:

$$\frac{\partial}{\partial p} \ln L_p(x_1, \ldots, x_n) \;=\; \frac{n\,\bar{x}}{p} - \frac{n\,(1-\bar{x})}{(1-p)} \overset{!}{=} 0.$$

We obtain $\hat{p} = \bar{x}$.

Since

$$\frac{\partial^2}{\partial p^2} \ln L_p(x_1, \ldots, x_n) \;=\; -\frac{n\,\bar{x}}{p^2} - \frac{n\,(1-\bar{x})}{(1-p)^2} \;<\; 0$$

the ML estimator for p is $\hat{\Theta} = \bar{X}$.

Example: Exponential distribution
Let $X_i \sim Exp(\lambda)$ for $i = 1, .., n$. Then

$$
\begin{aligned}
L_\lambda(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \lambda \exp(-\lambda\, x_i) = \lambda^n \exp(-\lambda\, n\, \bar{x}), \\
\ln L_\lambda(x_1, \ldots, x_n) &= n \ln \lambda - n\, \lambda\, \bar{x}.
\end{aligned}
$$

The solution is

$$\frac{d}{d\lambda}\big(\ln L_\lambda(x_1, \ldots, x_n)\big) = n/\lambda - n\, \bar{x} \overset{!}{=} 0$$

leading to $\hat{\lambda} = 1/\bar{x}$.

ML-estimators

Distribution	ϑ	ML-estimates
$\mathrm{B}(1;p)$	$p\ (=\mu)$	$\bar{X}$
$\mathrm{Exp}(\lambda)$	λ	$\frac{1}{\bar{X}}$
$\mathrm{P}(\lambda)$	$\lambda\ (=\mu=\sigma^2)$	$\bar{X}$
$\mathrm{N}(\mu;\sigma^2)$	μ	$\bar{X}$
$\mathrm{N}(\mu;\sigma^2)$	σ^2	$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$
$\mathrm{N}_k(\boldsymbol{\mu};\boldsymbol{\Sigma})$	$\boldsymbol{\mu}$	$\frac{1}{n}\boldsymbol{X}\mathbf{1}$
$\mathrm{N}_k(\boldsymbol{\mu};\boldsymbol{\Sigma})$	$\boldsymbol{\Sigma}$	$\frac{1}{n}\boldsymbol{X}(\boldsymbol{I}-\frac{1}{n}\mathbf{1}\mathbf{1}')\boldsymbol{X}'$

Bayes estimation

Idea: the true parameters are unknown and thus can be treated as realisations of some random variables.

- Given:
  - A random sample $(x_1, \ldots, x_n)$
  - Joint density (Likelihood function) $f(x_1, \ldots, x_n|\vartheta)$
  - Prior information about ϑ (assessment of the analyst) in form of a priori density $\varphi(\vartheta)$
- a posteriori density $\psi(\vartheta|x_1, \ldots, x_n)$

Recall the Bayes formula

$$P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{\sum\limits_i P(B|A_i) \cdot P(A_i)}$$

with

$$\begin{array}{lll} P(A_j|B) & \text{replaced by} & \psi(\vartheta|x_1,\ldots,x_n) \\ P(B|A_j) & \text{replaced by} & f(x_1,\ldots,x_n|\vartheta) \\ P(A_j) & \text{replaced by} & \varphi(\vartheta) \end{array}$$

$$\Rightarrow \psi(\vartheta|x_1,\ldots,x_n) = \begin{cases} \frac{f(x_1,\ldots,x_n|\vartheta_i)\cdot\varphi(\vartheta_i)}{\sum\limits_j f(x_1,\ldots,x_n|\vartheta_j)\cdot\varphi(\vartheta_j)} & \text{discrete case} \\ \dfrac{f(x_1,\ldots,x_n|\vartheta)\cdot\varphi(\vartheta)}{\int\limits_{-\infty}^{\infty} f(x_1,\ldots,x_n|\vartheta)\cdot\varphi(\vartheta)\,\mathrm{d}\vartheta} & \text{continuous case} \end{cases}$$

Bayes-estimation

1. fix the a priori density: $\varphi(\vartheta)$
2. Build the joint density : $f(x_1, \ldots, x_n|\vartheta)$
3. determine the posteriori density : $\psi(\vartheta|x_1, \ldots, x_n)$
4. For example, determine the location parameters of the posterior density
    - median,
    - mean, etc.

What is a “good” estimator?

Idea 1: the estimator should be close (?) to the true value.

- Bias: should be small (if zero $\rightsquigarrow$ unbiased)

$$Bias(\hat{\vartheta}) = E(\hat{\vartheta}) - \vartheta$$

- Variance: should be small $\rightsquigarrow$ efficiency

$$Var(\hat{\vartheta}) = E[\hat{\vartheta} - E(\hat{\vartheta})]^2$$

- MSE - mean-squared error: trade-off between bias and variance

$$MSE(\hat{\vartheta}) = [Bias(\hat{\vartheta})]^2 + Var(\hat{\vartheta})\}$$

Example:

Are

$$\hat{\Theta} = \bar{X}_n, \quad \hat{\Theta}' = \frac{X_1 + X_n}{2}, \quad \hat{\Theta}'' = \frac{1}{n-1}\sum_{i=1}^{n} X_i$$

unbiased for $E(X_i) = \mu$?

- $E(\hat{\Theta}) = E(\bar{X}_n) = \mu$
  $\Rightarrow$ unbiased
- $E(\hat{\Theta}') = E\left(\frac{X_1+X_n}{2}\right) = \frac{1}{2}[E(X_1) + E(X_n)] = \frac{1}{2}(\mu + \mu) = \mu$
  $\Rightarrow$ unbiased
- $E(\hat{\Theta}'') = E\left(\frac{1}{n-1}\sum_{i=1}^{n} X_i\right) = \frac{1}{n-1}\sum_{i=1}^{n} E(X_i) = \frac{1}{n-1}\sum_{i=1}^{n} \mu =$
  $\frac{n}{n-1} \cdot \mu \underset{n\to\infty}{\longrightarrow} \mu$
  $\Rightarrow$ asymptotically unbiased.

Which estimator better?

Efficiency

- Let $\hat{\Theta}$ be $\hat{\Theta}'$ two unbiased estimators of θ. Then $\hat{\Theta}$ is **more efficient** than $\hat{\Theta}'$, if independently of the numeric value of θ it holds:

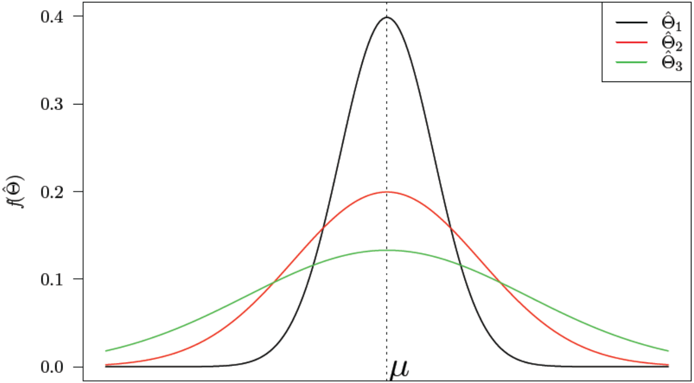
$$Var(\hat{\Theta}) < Var(\hat{\Theta}')$$

Example:

$$\begin{aligned} Var(\hat{\Theta}) = \quad & Var(\bar{X}) = \frac{\sigma^2}{n} < \\ < \quad & Var(\hat{\Theta}') = Var\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2} \end{aligned}$$

(if $n > 2$) $\hat{\Theta}$ is more efficient than $\hat{\Theta}'$.

An unbiased estimator with the smallest possible variance (within a particular model) is called the most efficient estimator.



Idea 2: what happens to the estimator if the sample size increases?
A sequence of estimators $\hat{\vartheta}_n$

$$\begin{aligned}
\hat{\Theta}_1 &= g_1(X_1) \\
\hat{\Theta}_2 &= g_2(X_1, X_2) \\
&\vdots \\
\hat{\Theta}_n &= g_n(X_1, \ldots, X_n)
\end{aligned}$$

is **consistent** for θ, if for all $c > 0$ it holds:

$$P(|\hat{\Theta}_n - \theta| \geq c) \underset{n\to\infty}{\longrightarrow} 0$$

Convergence in probability

A sequence of RV's z_n converges in probability to a constant c if for all $\varepsilon > 0$

$$P(|z_n - c| > \varepsilon) \longrightarrow 0, \text{ as } n \to \infty.$$

We write $z_n \xrightarrow{p} z$ and say that c is the probability limit of z_n, i.e. $p\lim z_n = c$.

Thus an estimator is consistent if

$$p \lim_{n\to\infty} \hat{\vartheta} = \vartheta$$

The weak low of large numbers

1. Tschebyscheff inequality

$$P(|X - E(X)| \geqq c) \leq \frac{Var(X)}{c^2}$$

2. applying to $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ results in

$$P(|\bar{X}_n - \mu| \geqq c) \leq \frac{\sigma^2}{n \cdot c^2}$$

3. as $n \to \infty \Rightarrow$ low of large numbers (LLN):

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq c) = 0 \qquad \text{or} \qquad \lim_{n\to\infty} P(|\bar{X}_n - \mu| \leq c) = 1$$

Checking the limit in the definition of consistency is not trivial. But from Tschebysheff inequality we have the following sufficient conditions for consistency:

$$\left(\lim_{n\to\infty}\right)E(\hat{\Theta}_n) = \theta \qquad \text{and} \qquad \lim_{n\to\infty} Var(\hat{\Theta}_n) = 0.$$

Example: Is $\bar{X}_n$ consistent for μ?
Es gilt:

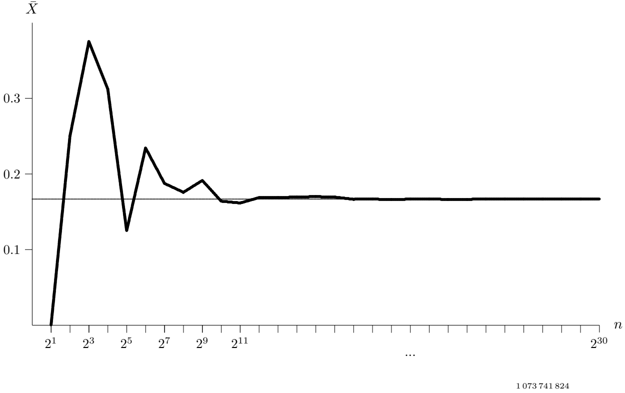
- $E(\bar{X}_n) = \mu,$ i.e. $\bar{X}_n$ is unbiased for μ.
- $Var(\bar{X}_n) = \frac{\sigma^2}{n} \underset{n\to\infty}{\longrightarrow} 0,$ i.e. the variance converges to zero.

Thus $\bar{X}_n$ is consistent for μ.

Example: toss a die

$$X_i = \begin{cases} 1 & \text{if you toss a 6} \\ 0 & \text{if you toss other number} \end{cases}$$

Since $X_i \sim B(1, 1/6)$ then $E(\bar{X}) = 1/6$ and
$Var(\bar{X}) = \frac{1}{n} \cdot Var(X_i) \; = \; \frac{5}{36\,n}$.



Thus: any estimator should be:

- Unbiased: $E(\hat{\vartheta}) = \vartheta$
- Efficient: $Var(\hat{\vartheta})$ is the smallest among all other unbiased estimators of ϑ
- or asymptotically efficient: $\lim_{n\to\infty} Var(\hat{\vartheta}) = [I(\vartheta)]^{-1}$ converges to the smallest possible variance given by the Cramer-Rao lower bound

$$I(\vartheta) = -E\left[\frac{\partial^2 \ell(x_1, ..., x_n|\vartheta)}{\partial\vartheta\partial\vartheta'}\right] = E\left[\frac{\partial \ell(x_1, ..., x_n|\vartheta)}{\partial\vartheta}\frac{\partial \ell(x_1, ..., x_n|\vartheta)}{\partial\vartheta'}\right]$$

- Consistent:

$$p\lim_{n\to\infty} \hat{\vartheta} = \vartheta$$

- Robust: $\hat{\vartheta}$ is still a good estimator even if the distributional assumptions are not satisfied.

Note: frequently is is important to know the distribution of $\hat{\vartheta}$ or at least its asymptotic distribution

Example: Let $\vartheta = \mu$ and $\hat{\vartheta} = \bar{x}$.

- If the sample is drawn from a normal distribution then

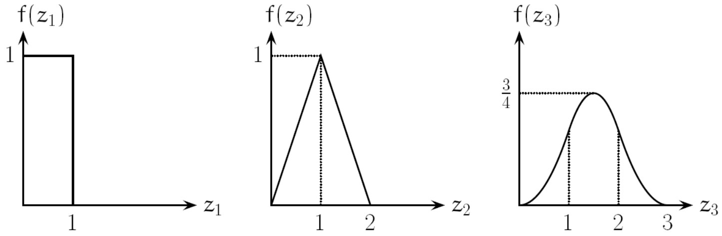
$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n).$$

- What happens if the sample is not normal? $\rightsquigarrow$ CLT

Central limit theorem (CLT)

The CLT gives a statement about the asymptotic behavior of the mean (or any simple function of the sample)

Example: X_1, X_2, X_3 in $[0;1]$ uniformly distributed;
$Z_1 = X_1,\;\; Z_2 = X_1 + X_2,\;\; Z_3 = X_1 + X_2 + X_3$



1. Standardisation:

$$Y_n = \frac{\bar{X}_n - \mu}{\sigma \cdot \frac{1}{\sqrt{n}}} = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$$

2.

Central Limit Theorem

$$\lim_{n\to\infty} P(Y_n \leqq x) = \lim_{n\to\infty} P\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq x\right) = \Phi(x)$$

3. We write

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$$

Example: wear parts of a machine

(a) A machine contains 100 wear parts. Each part should be replaced during the next year with the prob. of $\frac{1}{6}$. What is the prob. that we have to replace more than 10 but less than 21 parts?

$$P\left(10 < \sum_{i=1}^{100} X_i \leq 20\right) = F_{B(100,\frac{1}{6})}(20) - F_{B(100,\frac{1}{6})}(10)$$

$$= \sum_{i=11}^{20} \binom{100}{i} \left(\frac{1}{6}\right)^i \left(\frac{5}{6}\right)^{100-i}$$

Problem: $F_{B(100,\frac{1}{6})}(x)$ for $1 < x < 100$!

(b) Approximation with CLT $(n = 100, E(X_i) = \frac{1}{6}, Var(X_i) = \frac{5}{36})$

$$
\begin{aligned}
P\left(10 < \sum_{i=1}^{100} X_i \leq 20\right) &= P\left(0.1 < \bar{X} \leq 0.2\right) \\
&= P\left(\underbrace{\sqrt{100}\,\frac{0.1-\frac{1}{6}}{\sqrt{\frac{5}{36}}}}_{=-1.7888} < \sqrt{100}\,\frac{\bar{X}-\frac{1}{6}}{\sqrt{\frac{5}{36}}} \leq \underbrace{\sqrt{100}\,\frac{0.2-\frac{1}{6}}{\sqrt{\frac{5}{36}}}}_{=0.8944}\right) \\
&\approx \Phi(0.8944) - \Phi(-1.7888) \\
&= \Phi(0.8944) - 1 + \Phi(1.7888) \\
&= 0.8143 - 1 + 0.9631 \\
&= 0.7774
\end{aligned}
$$

Theorem (Multivariate Central Limit Theorem)

If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *are a random sample from an arbitrary multivariate distribution with finite mean* $\boldsymbol{\mu}$ *and positive definite covariance matrix* $\boldsymbol{\Sigma}$*, then* $\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Theorem (Multivariate Central Limit Theorem with unequal moments (Lindberg-Feller))

If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *are a set of RVs with finite means* $\boldsymbol{\mu}_i$ *and finite positive definite covariance matrices* $\boldsymbol{\Sigma}_i$*. Let* $\bar{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\mu}_i$ *and* $\bar{\boldsymbol{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Sigma}_i$*. If no single term dominates the average variance, i.e.* $\lim_{n\to\infty}(n\bar{\boldsymbol{\Sigma}}_n)^{-1}\boldsymbol{\Sigma}_i = \boldsymbol{O}$*, and if the average variance converges to a finite constant* $\bar{\boldsymbol{\Sigma}} = \lim_{n\to\infty} \bar{\boldsymbol{\Sigma}}_n^2$*, then*

$$\sqrt{n}(\bar{\boldsymbol{X}}_n - \bar{\boldsymbol{\mu}}_n) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \bar{\boldsymbol{\Sigma}}).$$

Theorem (Limiting normal distribution of a function (Delta-method))

If $\sqrt{n}(z_n - \mu) \xrightarrow{d} M(0, \sigma^2)$ *and if* $g(\cdot)$ *is a continuous function not involving* n, *then*

$$\sqrt{n}(g(z_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, [g'(\mu)]^2\sigma^2).$$

Theorem (Limiting normal distribution of a multivariate function (Delta-method))

If $\boldsymbol{z}_n$ *is a sequence of* $k \times 1$*-dimensional vector-valued RVs such that* $\sqrt{n}(\boldsymbol{z}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ *and if* $\boldsymbol{c}(\cdot) : \mathbb{R}^k \to \mathbb{R}^J$ *is a continuous function not involving* n, *then*

$$\sqrt{n}(\boldsymbol{c}(\boldsymbol{z}_n) - \boldsymbol{c}(\boldsymbol{\mu})) \xrightarrow{d} \mathcal{N}_J(\mathbf{0}, \boldsymbol{C}(\boldsymbol{\mu})\boldsymbol{\Sigma}\boldsymbol{C}(\boldsymbol{\mu})')$$

where $\boldsymbol{C}(\boldsymbol{\mu})$ *is the* $J \times k$ *matrix of first partial derivatives* $\partial\boldsymbol{c}(\boldsymbol{\mu})/\partial\boldsymbol{\mu}'$.

Confidence intervals

Assumption: it holds $X_i \sim F_\vartheta$, $\vartheta \in \Theta$, where $X_1, \ldots, X_n$ are the sample variables

Aim: provide an area (e. g. interval), where the unknown parameter ϑ will belong to with a high probability

Let $T_1(\boldsymbol{X})$, $T_2(\boldsymbol{X})$ be functions of the sample with $T_1 \leq T_2$ and $\alpha \in (0, 1)$.
The interval $\big[T_1(\boldsymbol{X}), T_2(\boldsymbol{X})\big]$ with

$$P_\vartheta(T_1 \leq \vartheta \leq T_2) \quad = \quad 1 - \alpha \qquad \forall \vartheta \in \Theta \qquad\qquad (*)$$

is an exact (two-sided) confidence interval (CI) for ϑ with the confidence level $1 - \alpha$.

Note: In practice we select α often equal to 0.1, 0.05 or 0.01.

Interpretation: If $[T_1, T_2]$ is a 90%-CI, then the unknown parameter ϑ belongs to this interval with the probability of 90%.

$\big[T_1(\boldsymbol{X}), \infty\big)$ is a one-sided lower confidence interval at the confidence level $1-\alpha$, if

$$P_\vartheta\big(T_1 \leq \vartheta\big) \;=\; 1-\alpha \quad \forall\,\vartheta \in \Theta$$

and $\big(-\infty, T_2(\boldsymbol{X})\big]$ is a one-sided upper confidence interval at the confidence level $1-\alpha$, if

$$P_\vartheta\big(\vartheta \leq T_2\big) \;=\; 1-\alpha \quad \forall\,\vartheta \in \Theta$$

Example:

- Assume that the parameter measures the riskiness of an asset. Then we are interested in the upper confidence interval, since ϑ should be bounded from above.
- If ϑ denotes the tear strength of a rope, then we consider the lower confidence interval, since the lower bound cannot be undershot.

Construction of confidence intervals

Confidence interval for normally distributed random variables

Assumption: The sample variables $X_1, ..., X_n$ are independent and normally distributed with $X_i \sim N(\mu, \sigma^2)$ for $i = 1, .., n$.

i) Confidence interval for μ (σ is known)

Starting point: estimate μ with $\bar{X}$

Since $Var(\bar{X}) = \sigma^2/n$, we postulate the following structure of the confidence interval

$$[\bar{X} + c_1 \frac{\sigma}{\sqrt{n}}, \bar{X} + c_2 \frac{\sigma}{\sqrt{n}}]$$

with suitable constants $c_1 < 0 < c_2$. c_1 and c_2 will depend on α, i.e. of (*).

Since $\bar{X} \sim N(\mu, \sigma^2/n)$, i.e. the distribution is symmetric around μ, we select $c_2 = c = -c_1$.

Thus

$$\mu \in \left[\bar{X} - c\,\frac{\sigma}{\sqrt{n}}, \bar{X} + c\,\frac{\sigma}{\sqrt{n}}\right] \;\Leftrightarrow\; \sqrt{n}\,\frac{|\bar{X}-\mu|}{\sigma} \le c$$

Since $\sqrt{n}(\bar{X}-\mu)/\sigma \sim \Phi$, we determine c in such way, that

$$\begin{aligned}
&P_\mu\left(\sqrt{n}\,\frac{|\bar{X}-\mu|}{\sigma} \le c\right) \\
&= P_\mu\left(\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \le c\right) - P_\mu\left(\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} < -c\right) \\
&= \Phi(c) - \Phi(-c) \;=\; 2\,\Phi(c) - 1 \overset{!}{=} 1-\alpha\,.
\end{aligned}$$

This implies $c = z_{1-\alpha/2}$

The exact CI for μ at the level $1-\alpha$

$$\left[\bar{X} - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right]$$

Example: demand for a particular product

A manufacturer had on average $\bar{X} = 100$ orders per day for a particular product during the last $n = 100$ days. From experience they know that the standard deviation σ is 100. To satisfy the demand, the company is interested in the confidence interval for the expected demand.

Assuming normally distributed random variables and taking $\alpha = 0.05$ we obtain $z_{0.975} = 1.96$ and

$$\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 100 \pm 1.96\frac{100}{\sqrt{100}} = 100 \pm 19.6$$

The 95%-CI for μ is given by $[80.4, 119.6]$.

ii) Confidence interval for μ (σ is unknown)

To derive the CI we use the same procedure as above. We estimate the σ^2 by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

The derivation is more complicated, since we have to determine the distribution of

$$\sqrt{n}\,\frac{\bar{X}-\mu}{S} \;\sim\; t_{n-1}.$$

Exact confidence interval for μ (σ is unknown) at the level $1-\alpha$

$$\left[\bar{X} - t_{n-1;1-\alpha/2}\,\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1;1-\alpha/2}\,\frac{S}{\sqrt{n}}\right]$$

Example: length of bolts in cm

The average length of $n = 125$ bolts is $\bar{x} = 10.25$. Moreover, $s = 0.3$. For $\alpha = 0.1$ we obtain $t_{124;0.95} = 1.6576$, and

$$\left[10.25 - 1.6576 \cdot \frac{0.3}{\sqrt{125}}, 10.25 + 1.6576 \cdot \frac{0.3}{\sqrt{125}}\right]$$
$$\approx \; [10.206, 10.294]$$

iii) Confidence intervals for σ^2

Estimator for σ^2: $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

It can be shown that

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

χ^2_{n-1} denotes the (central) χ^2–distribution with $n-1$ degrees of freedom.

Since the χ^2–distribution is not symmetric, the CI for σ^2 is also not symmetric. We take

$$P_\sigma\left(c_1 \le \frac{(n-1)\,S^2}{\sigma^2} \le c_2\right) = \chi^2_{n-1}(c_2) - \chi^2_{n-1}(c_1) \overset{!}{=} 1-\alpha$$

Since the equation contains two independent quantities, it cannot be solved uniquely. For this reason we assume

$$\chi^2_{n-1}(c_2) = 1 - \frac{\alpha}{2}, \quad \chi^2_{n-1}(c_1) = \frac{\alpha}{2}\ .$$

We use the notation $c_2 = \chi^2_{n-1;1-\alpha/2}$ and $c_1 = \chi^2_{n-1;\alpha/2}$.

Since

$$c_1 \leq \frac{(n-1)\,S^2}{\sigma^2} \leq c_2 \quad \Leftrightarrow \quad \frac{(n-1)\,S^2}{c_2} \leq \sigma^2 \leq \frac{(n-1)\,S^2}{c_1},$$

the exact confidence interval for σ^2 at the level $1-\alpha$ is given by

$$\left[\frac{(n-1)\,S^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)\,S^2}{\chi^2_{n-1;\alpha/2}}\right]$$

Example: the length of bolts in cm

Let $\alpha = 0.1$, $n = 100$, $s^2 = 0.1$.
Since $\chi^2_{99;0.05} = 77.93$ and $\chi^2_{99;0.95} = 124.3$, the CI is given by

$$\left[\frac{99 \cdot 0.1}{124.3}, \frac{99 \cdot 0.1}{77.93}\right]$$

$$\approx \; [0.08, 0.13]$$

now: Asymptotic confidence intervals

Using the the CLT we can derive the asymptotic CI. It holds

$$\lim_{n\to\infty} P_\vartheta(T_1(\boldsymbol{X}) \leq \vartheta \leq T_2(\boldsymbol{X})) = 1 - \alpha \quad \forall\, \vartheta \in \Theta.$$

Example: Let $X_1, ..., X_n$ be iid with $E(X_i) = \mu$ for all $i \geq 1$. The objective is a CI for μ.

i) $\sigma^2 = Var(X_i)$ is known

The CI has the same structure as in the case of normal distribution

$$\left[\bar{X} - c\,\frac{\sigma}{\sqrt{n}}, \bar{X} + c\,\frac{\sigma}{\sqrt{n}}\right]$$

Problem: The exact distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma$ is unknown.

For this reason one applies the CLT.

$$\lim_{n\to\infty} P_\mu(\mu \in \left[\bar{X} - c\,\frac{\sigma}{\sqrt{n}}, \bar{X} + c\,\frac{\sigma}{\sqrt{n}}\right]) = \lim_{n\to\infty} P_\mu\left(\sqrt{n}\,\frac{|\bar{X}-\mu|}{\sigma} \le c\right)$$

$$= \lim_{n\to\infty} P_\mu\left(\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \le c\right) - \lim_{n\to\infty} P_\mu\left(\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} < -c\right)$$

$$= \Phi(c) - \Phi(-c) = 2\,\Phi(c) - 1 \stackrel{!}{=} 1-\alpha\,.$$

Consequently $c = z_{1-\alpha/2}$ and an asymptotic CI for μ is given by

$$\left[\bar{X} - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right]$$

Rule of thumb: $n \ge 30$

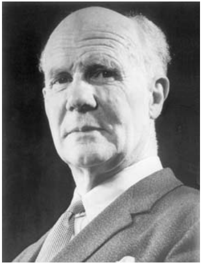
Statistical Tests

Basics: Jerzy Neyman (1894–1981) and Egon S. Pearson (1895–1980)

Example: is the filling capacity of bottles exactly 500ml as argued?

- It is a decision problem between two hypotheses.
- We speak of a test problem.
- These hypotheses can be expressed using the parameter of interest (here μ).

Jerzy Neyman (1894–1981) Egon S. Pearson (1895–1980)



Let $X \sim F_\vartheta$, $\vartheta \in \Theta$. Θ_0 and Θ_1 are disjunct and $\Theta_0 \cup \Theta_1 = \Theta$.

The question regarding ϑ is expressed in terms two hypotheses H_0 (null hypothesis) and H_1 alternative hypothesis:

$$H_0: \ \vartheta \in \Theta_0 \quad \text{versus} \quad H_1: \ \vartheta \in \Theta_1 \ .$$

Action: On the basis of the sample $x_1, \ldots, x_n$ we should make a decision whether if $\mu = 500\, ml$ is fulfilled or not.

Such a decision rule is called a statistical test.

Here we shortly sketch how does a test look like:

Example: filling capacity of bottles

$$H_0 : \mu = 500ml \quad \text{vs} \quad H_1 : \mu \neq 500ml.$$

Starting point: Estimator for μ, here $\hat{\mu} = \bar{x}$

Consideration

Reject H_0, when $|\bar{x} - 500|$ is large enough , e. g. when $|\bar{x} - 500| > c$ for some constant $c > 0 \rightsquigarrow$ i.e. reject H_0 if we have sufficient evidence against H_0.

Decision rule

$$\begin{aligned} |\bar{x} - 500| > c \quad &\rightsquigarrow \quad \text{reject } H_0 \text{ (accept } H_1\text{)}, \\ |\bar{x} - 500| \leq c \quad &\rightsquigarrow \quad \text{do not reject } H_0 \text{ (} H_0 \text{ CANNOT be accpeted).} \end{aligned}$$

Note: If the hypothesis is an inequality, then it must be chosen as H_1. A hypothesis with equality is always in H_0.

One sample Z-test

- We assume
  - $G \sim \mathrm{N}(\mu;\sigma)$ with known σ
  - random sample $X_1,\ldots,X_n$
- Different pairs of hypotheses

two-sided

$$H_0 \; : \; \mu = \mu_0 \qquad\qquad H_1 \; : \; \mu \neq \mu_0$$

one-sided test

right-sided test	$H_0 \; : \; \mu \leq \mu_0$	$H_1 \; : \; \mu > \mu_0$
left-sided test	$H_0 \; : \; \mu \geq \mu_0$	$H_1 \; : \; \mu < \mu_0$

- $X_1,\ldots,X_{25}$ with $X_i =$ capacity of the i-th bottle $\sim \mathrm{N}(\mu;1.5^2)$
  null hypotheses $H_0 : \; \mu = 500,$ i.e. $\mu_0 = 500$

Possible errors: rejection of H_0, if H_0 is correct and non-rejection of H_0, if H_0 is wrong:

Decision	**Reality**	
	rain	no rain
take umbrella	correct decision	error
do not take umbrella	error	correct decision

$\{H_1\}$: decide to reject H_0
$\{H_0\}$: decide not to reject H_0

decision	**Reality**	
	H_0 is correct	H_0 is not correct (H_1 is correct)
H_0 is not rejected: $\{H_0\}$	correct decision $\{H_0\}\|H_0$ $P(\{H_0\}\|H_0) = 1-\alpha$	error of the 2nd type $\{H_0\}\|H_1$ $P(\{H_0\}\|H_1) = \beta$
H_0 is rejected: $\{H_1\}$	error of the 1st type $\{H_1\}\|H_0$ $P(\{H_1\}\|H_0) = \alpha$	correct decision $\{H_1\}\|H_1$ $P(\{H_1\}\|H_1) = 1-\beta$

Level of significance α:
the highest allowed probability of the type 1 error.

Test statistic (test function)

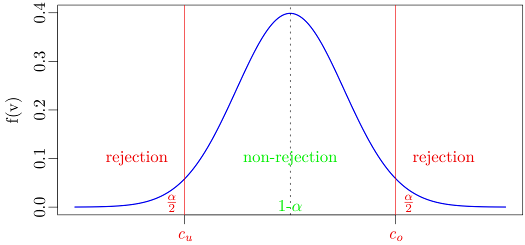
- for testing pusposes we aggregate the information in the sample to a test statistic
$$V = V(X_1, ..., X_n).$$
- The functional form of V depends on the test/hypotheses, ect.
- The distribution of V under H_0 should be known (at least asymptotically), i.e. if H_0 is correct
$$F(v|\mu_0)$$
- Using this distribution split the set of possible values of the test statistics into
  - rejection area - if v takes a value here then H_0 is rejected
  - non-rejection area - if v takes a value here then H_0 is NOT rejected

Critical values

The rejection area (or critical values) is determined in such was, that the probability of getting a test statistic in this area (assuming that H_0 is correct) is not higher then the given α

$$P(V \in \text{ rejection area of } H_0 \mid \mu_0) \leq \alpha$$

$$P(V \in \text{ non-rejection area of } H_0 \mid \mu_0) \geq 1 - \alpha$$



- two-sided test

$$H_0 \ : \ \mu = \mu_0 \qquad H_1 \ : \ \mu \neq \mu_0$$

- non-rejection area $[c_u; c_o]$

$$P(c_u \leq V \leq c_o \mid \mu_0) = 1 - \alpha$$

- rejection area $B = (-\infty; c_u) \cup (c_0; \infty)$

$$P(V \leq c_u \mid \mu_0) + P(V \geq c_o \mid \mu_0) = \alpha/2 + \alpha/2 = \alpha$$

- test statistic

$$V = \sqrt{n}\frac{|\bar{x} - \mu_0|}{\sigma}$$

- Relying on the symmetry of the normal distribution:

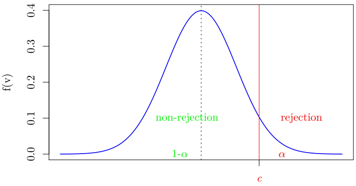
$$\begin{aligned} P(|V| > c) &= P(V > c \text{ or } -V > c) = P(V > c \text{ or } V < -c) \\ &= P(V > c) + P(V < -c) = 2 \cdot P(V > c) \\ &= 2 \cdot [1 - P(V \leq c)] = 2 \cdot [1 - \Phi(c)] \overset{!}{=} \alpha \iff \\ \Phi(c) &= 1 - \tfrac{\alpha}{2} \iff c = z_{1-\frac{\alpha}{2}} \end{aligned}$$

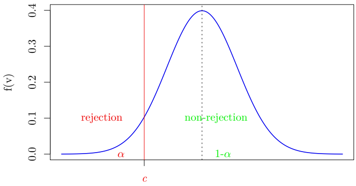
  H_0 is rejected if $|v| > z_{1-\frac{\alpha}{2}}$.
- $B = (-\infty; -z_{1-\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}; \infty)$ is the rejection area.

Similarly for one sided tests

H_0 is rejected if $v \in B$ with

$$\left.\begin{array}{llr} B &= \left(-\infty; -z_{1-\frac{\alpha}{2}}\right) \cup \left(z_{1-\frac{\alpha}{2}}; \infty\right) & \text{case a)} \\ B &= \left(-\infty; -z_{1-\alpha}\right) & \text{case b)} \\ B &= \left(z_{1-\alpha}; \infty\right) & \text{case c)} \end{array}\right\} \quad (1)$$





Example

$X_1, \dots, X_{25}$ with $X_i \sim \mathrm{N}(\mu; 1.5^2)$ and $\bar{x} = 499.28$
Test $H_0 : \mu = 500, H_1 : \mu \neq 500$ with $\alpha = 0.01$

1. $\alpha = 0.01$
2. $v = \frac{499.28-500}{1.5} \cdot \sqrt{25} = -2.4$
3. $\mathrm{N}(0;1): \; z_{1-\frac{\alpha}{2}} = z_{1-0.005} = z_{0.995} = 2.576$
   $\Rightarrow B = (-\infty; -2.576) \cup (2.576; \infty)$
4. $v \notin B \Rightarrow H_0$ not rejected

$\rightsquigarrow$ With significance level of 1% we cannot prove that the capacity deviates from the stated capacity.

p-value

- If you change α you have to run the test again $\rightsquigarrow$ stat software do not ask for α, but compute the p-value which allows you to run the test for any α
- p-value is the largest level of significance for which H_0 is still not rejected
- The smaller the p-value is, the more evidence the sample contains against H_0
- decision rule

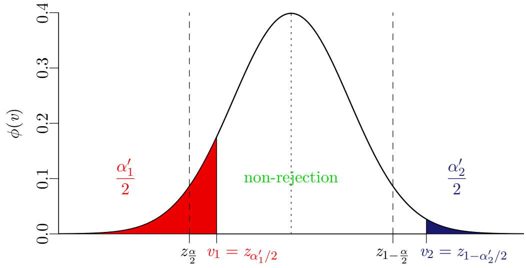
$$\boxed{\begin{array}{ll} \alpha > \alpha' & \Rightarrow H_0 \text{ rejected} \\ \alpha \leq \alpha' & \Rightarrow H_0 \text{ not rejected} \end{array}}$$

Case a: two-sided Z-test

The smallest value of α, for which H_0 is still not rejected, satisfies

$$\begin{cases} -z_{1-\frac{\alpha'}{2}} = z_{\frac{\alpha'}{2}} = v, & \text{if } v < 0 \\ z_{1-\frac{\alpha'}{2}} = -z_{\frac{\alpha'}{2}} = v, & \text{if } v > 0 \end{cases}$$

We are looking for such a value of α', that $\Phi(v) = \frac{\alpha'}{2}$ (if $v < 0$) or $\Phi(v) = 1 - \frac{\alpha'}{2}$ (if $v > 0$).



Example

Using the quantiles of $\mathcal{N}(0,1)$ we obtain

$$\Phi(-2.4) = 1 - \Phi(2.4) = 1 - 0.9918 = 0.0082 = \tfrac{\alpha'}{2}.$$

Thus $\alpha' = 0.0164$

Let $\alpha = 0.01$. Since $\alpha' > \alpha$, we cannot reject H_0 at $\alpha = 0.01$.
Let $\alpha = 0.05$. Since $\alpha' < \alpha$, we reject H_0 at $\alpha = 0.05$.

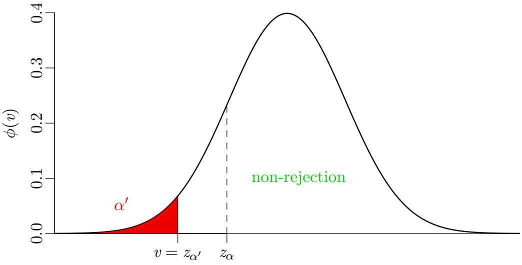
$\rightsquigarrow$ For all $\alpha' < \alpha$ we can reject H_0.

Case b: left-sided Z-test

The smaller value of α, for which H_0 is still not rejected, satisfies

$$-z_{1-\alpha'} = z_{\alpha'} = v$$

We are looking for such value of α', that $\Phi(v) = \alpha'$.



Note:

In Case a) we can also run the test using the confidence intervals: We compute for given α the symmetric CI $[v_u; v_o]$ centered at $\bar{x}$ and reject $H_0: \ \mu = \mu_0$ vs. $H_1: \ \mu \neq \mu_0$ if $\mu_0 \notin [v_u; v_o]$.

Here: CI for μ :

1. $1 - \alpha = 1 - 0.01 = 0.99$
2. $\mathrm{N}(0;1): \ c = z_{1-\frac{\alpha}{2}} = z_{1-\frac{0.01}{2}} = z_{0.995} = 2.576$
3. $\bar{x} = 499.28$
4. $\frac{\sigma c}{\sqrt{n}} = \frac{1.5 \cdot 2.576}{\sqrt{25}} = 0.77$
5. $[499.28 - 0.77; 499.28 + 0.77] = [498.51; 500.05]$

$\mu_0 = 500 \in [498.51; 500.05] \Rightarrow H_0$ cannot be rejected

Other tests

- Test of $H_0 : \mu = \mu_0$ with $X_i \sim N(\mu, \sigma^2)$, but σ^2 is unknown
  - Estimate σ^2 by s^2
  - $V = \sqrt{n}\frac{\bar{X}-\mu_0}{s} \sim t_{n-1}$
  - The rejection area for a two-sided test is

$$B = (-\infty; -t_{n-1;1-\alpha/2}) \cup (t_{n-1;1-\alpha/2}; +\infty)$$

- Test of $H_0 : \mu = \mu_0$ if the distribution is unknown (asymptotic Z-test)
  - Rely on the CLT
  - $V = \sqrt{n}\frac{\bar{X}-\mu_0}{s} \overset{approx}{\sim} \mathcal{N}(0,1)$
  - The rejection areas as for the simple Z-test

Other tests

- Test of $H_0 : p = p_0$, with $X_i \sim B(n, p)$
  - Check if $5 \leq \sum x_i \leq n - 5$
  - Estimate $\hat{p} = \bar{x}$
  - Compute the test statistic $v = \sqrt{n}\frac{\bar{x}-p_0}{\sqrt{p_0(1-p_0)}} \overset{asymp}{\sim} \mathcal{N}(0,1)$
  - Follow the idea of the asymptotic Z-test
- Test of $H_0 : \sigma^2 = \sigma_0^2$, with $X_i \sim N(\mu, \sigma^2)$
  - Estimate σ^2 by s^2
  - $V = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$
  - The rejection area for a two-sided test is

$$B = (0; \chi^2_{n-1;\alpha/2}) \cup (\chi^2_{n-1;1-\alpha/2}; +\infty)$$

Example

$X_1, \dots, X_{2000} \sim \mathrm{B}(1; p)$ mit

$$X_i = \begin{cases} 1, & \text{falls } i\text{th person voted for the party A} \\ 0, & \text{else} \end{cases}$$

$$\sum_{i=1}^{2000} x_i = 108$$

Test $H_0: \; p \leq 0.05$ vs. $H_1: \; p > 0.05$ with $\alpha = 2\,\%$

Asymptotic Z-test, Case (c); $5 \leq \sum x_i \leq n - 5$: $\; 5 \leq 108 \leq 2000 - 5$

1. $\alpha = 0.02$
2. $v = \frac{\frac{108}{2000} - 0.05}{\sqrt{0.05 \cdot (1 - 0.05)}} \sqrt{2000} = 0.82$
3. $\mathrm{N}(0; 1): \; z_{1-\alpha} = z_{0.98} = 2.05 \Rightarrow B = (2.05; \infty)$
4. $v \notin B \Rightarrow H_0$ not rejected

Two-sample tests

- Given
  - two independent samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ with
  - sample sizes n_1 and.n_2
  - expectations $E(X_i) = \mu_1$ and $E(Y_i) = \mu_2$
  - variances $Var(X_i) = \sigma_1^2$ and $Var(Y_i) = \sigma_2^2$
  - means $\bar{X}$ and $\bar{Y}$
  - sample variances S_1^2 and S_2^2
- Object of interest:
  - Comparison of means/expectations $\mu_1 \lesseqqgtr \mu_2$
  - Comparison of variances $\sigma_1^2 \lesseqqgtr \sigma_2^2$

Comparison of means

Hypotheses

$$\boxed{\begin{array}{lll} \text{a)} & H_0: \; \mu_1 = \mu_2 & H_1: \; \mu_1 \neq \mu_2 \\ \text{b)} & H_0: \; \mu_1 \geq \mu_2 & H_1: \; \mu_1 < \mu_2 \\ \text{c)} & H_0: \; \mu_1 \leq \mu_2 & H_1: \; \mu_1 > \mu_2 \end{array}} \tag{2}$$

Estimator for $\mu_1 - \mu_2$: $\quad \bar{X} - \bar{Y}$

Two-sample Z-test

If the variance σ_1^2 and σ_2^2 are known, then

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Thus the test statistics is

$$V = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Under H_0 $(\mu_1 = \mu_2)$ and for Gaussian samples it holds

$$V \sim N(0, 1).$$

Two-sample t-test

If the variances σ_1^2 and σ_2^2 are unknown, but $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then estimate σ^2 with

$$\tilde{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

with (under H_0)

$$\frac{(n_1 + n_2 - 2)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2).$$

The test statistic is

$$V = \frac{\bar{X} - \bar{Y}}{\sqrt{\tilde{\sigma}^2 \frac{n_1 + n_2}{n_1 n_2}}}.$$

Under H_0 $(\mu_1 = \mu_2)$ it holds

$$V \sim t_{n_1 + n_2 - 2}.$$

Asymptotic two-sample Z-test

If the variances σ_1^2 and σ_2^2 are unknown and arbitrary, then the test statistic is

$$V = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Under H_0 ($\mu_1 = \mu_2$) and from the CLT it holds

$$V \overset{\text{approx.}}{\sim} N(0,1).$$

The rejection area is given in all three situation by

$$B = (-\infty; -x_{1-\frac{\alpha}{2}}) \cup (x_{1-\frac{\alpha}{2}}; \infty) \qquad \text{in case a)}$$

$$B = (-\infty; -x_{1-\alpha}) \qquad \text{in case b)}$$

$$B = (x_{1-\alpha}; \infty) \qquad \text{in case c)}$$

with the corresponding quantiles defined by the above distributions of the test statistics.

	Assumption	test statistics V	Distr. of V under H_0
1.	$X_i \sim N(\mu_1;\sigma_1^2)$ $Y_i \sim N(\mu_2;\sigma_2^2)$ σ_1^2 and σ_2^2 known	$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$V \sim N(0;1)$
2.	$X_i \sim N(\mu_1;\sigma_1^2)$ $Y_i \sim N(\mu_2;\sigma_2^2)$ σ_1^2 and σ_2^2 unknown but $\sigma_1^2 = \sigma_2^2$	$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\cdot\frac{n_1+n_2}{n_1 n_2}}}$	$V \sim t_{n_1+n_2-2}$
3.	$X_i \sim N(\mu_1;\sigma_1^2)$ $Y_i \sim N(\mu_2;\sigma_2^2)$ σ_1^2 and σ_2^2 unknown, but $\sigma_1^2 \neq \sigma_2^2$, $n_1 = n_2 = n$	$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\cdot\frac{n_1+n_2}{n_1 n_2}}}$	$V \overset{\text{approx.}}{\sim} t_{(n-1)\left[1+\frac{2}{s_1^2/s_2^2+s_2^2/s_1^2}\right]}$
4.	$X_i \sim B(1;p_1)$ $Y_i \sim B(1;p_2)$ $5 \leq \sum x_i \leq n_1 - 5$ $5 \leq \sum y_i \leq n_2 - 5$	$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(\sum X_i+\sum Y_i)(n_1+n_2-\sum X_i-\sum Y_i)}{(n_1+n_2)n_1 n_2}}}$	$V \overset{\text{approx.}}{\sim} N(0;1)$
5.	X_i, Y_i arbitr. distr. $n_1 > 30;\ n_2 > 30$ σ_1^2, σ_2^2 unknown	$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$	$V \overset{\text{approx.}}{\sim} N(0;1)$
6.	X_i, Y_i arbitr. distr. $n_1 > 30;\ n_2 > 30$ σ_1^2, σ_2^2 unknown	$\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$V \overset{\text{approx.}}{\sim} N(0;1)$

1. two-sample Z-test; 2./3. two-sample t-test;
4./5./6. approx. Z-test

Example: has the expected return of an asset increased after the announcement of the acquisition?

Let X_1 be the return before the announcement and X_2 the return after. Assume $X_i \sim N(\mu_i, \sigma_i)$ and X_1 and X_2 are independent. D

$$H_0 : \mu_1 \geq \mu_2 \qquad vs \qquad H_1 : \mu_1 < \mu_2$$

$n_1 = 115$, $\bar{x}_1 = 6.5$, $s_1 = 0.4$, $n_2 = 110$, $\bar{x}_2 = 8.14$ and $s_2 = 0.78$. Thus

$$v = \frac{6.5 - 8.14}{\sqrt{\frac{0.4^2}{115} + \frac{0.78^2}{110}}} = -19.71.$$

We obtain $z_{0.99} = 2.3263$ and $B = (-\infty; -2.3263)$. Since $v < -2.3263$ we reject H_0 and conclude that the expected return is significantly larger after the announcement.

Example:

$$X_1, \ldots, X_{80} \sim \mathrm{B}(1; p_1)$$

with

$$X_i = \begin{cases} 1, & \text{if the } i\text{th product is defective} \\ 0, & \text{else} \end{cases}, \qquad \sum_{i=1}^{80} x_i = 20$$

$$Y_1, \ldots, Y_{100} \sim \mathrm{B}(1; p_2)$$

with

$$Y_i = \begin{cases} 1, & \text{if the } i\text{th product is defective} \\ 0, & \text{else} \end{cases}, \qquad \sum_{i=1}^{100} y_i = 50$$

Can we argue that the probability of being defective is higher for Type 1 products than for Type 2 products?

1. $\alpha = 0.1$
2. $\bar{x} = \frac{20}{80} = 0.25;\quad \bar{y} = \frac{50}{100} = 0.5;\quad v = \frac{0.25-0.5}{\sqrt{\frac{(20+50)(80+100-20-50)}{(80+100)\cdot 80\cdot 100}}} = -3.42$
3. $\mathrm{N}(0;1):\quad z_{1-\alpha} = z_{0.9} = 1.282 \Rightarrow B = (-\infty; -1.282)$
4. $v \in B \Rightarrow H_0$ rejected, i.e. $p_1 < p_2$ is confirmed

Test for correlation/dependence

Assumption: let (X, Y) follow a 2–dim. normal distribution

$$E(X) = \mu_x \; , \; Var(X) \; = \; \sigma_x^2 \, ,$$

$$E(Y) = \mu_y \; , \; Var(Y) \; = \; \sigma_y^2 \, .$$

Let

$$\rho = Corr(X, Y) \; := \; \frac{Cov(X, Y)}{\sigma_x \, \sigma_y} \; = \; \frac{E\big((X - \mu_x)\,(Y - \mu_y)\big)}{\sigma_x \, \sigma_y} \, .$$

Note: if the samples are normal, then zero correlation implies independence.

The estimator for ρ is

$$r_{XY} = \hat{\rho} = \frac{s_{XY}}{s_X\, s_Y} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2 \sum\limits_{i=1}^{n}(y_i - \bar{y})^2}}$$

Note: similar tests can be derived the contingency tables and for the rank correlation of Spearman.

$$r_{SP} = \frac{\sum_{i=1}^{n}\left(R(x_i) - \bar{R}\right)\left(R(y_i) - \bar{R}\right)}{\sqrt{\sum_{i=1}^{n}\left(R(x_i) - \bar{R}\right)^2 \sum_{i=1}^{n}\left(R(y_i) - \bar{R}\right)^2}}$$

with $\bar{R} = \frac{n+1}{2}$

Test problem

- $H_0 : \rho = 0$ (i.e. X and Y are uncorrelated / independent assuming normality) vs
- $H_1 : \rho \neq 0$ (i.e. X and Y are dependent)

- Test statistics: $v = \sqrt{n-2}\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}$.
- Under H_0 it holds $V \overset{approx}{\sim} \mathcal{N}(0,1)$ (or t_{n-2} for small samples)
- Rejection area

$$B = (-\infty; -t_{n-2;1-\frac{\alpha}{2}}) \cup (t_{n-2;1-\frac{\alpha}{2}}; \infty)$$

Note: Similarly with r_{SP} with “no monotone dependence”

Example: test with $\alpha = 0.05$ if the is a significant correlation between the body height of fathers (Y) and sons at the age of 5 (X)?

x_i	109	114	116	105	114	116	114	108	108	122	117	115
115	112	122	113	108	109	115	108	118	110	113	111	116
y_i	167	176	186	175	175	182	180	172	185	186	183	178
175	175	180	181	172	179	170	172	172	172	176	180	182

It holds $n = 25$, $\bar{x} = 113.12$, $s_X = 4.352$, $\bar{y} = 177.24$, $s_Y = 5.206$ and $s_{XY} = 10.05333$. Thus $\hat{\rho} = 0.44365$ and

$$v = \sqrt{23}\,\frac{0.44365}{\sqrt{1 - 0.44365^2}} = 2.374\,.$$

For $\alpha = 0.05$ it holds $t_{23;0.975} = 2.069$ and

$$B = (-\infty; -2.069) \cup (2.069; +\infty)\,.$$

Since $v \in B$, we conclude that $H_0 : \rho = 0$ can be rejected.

Kolmogorov-Smirnov Goodness-of-Fit Test

Requirement: an independent random sample $X_1, \ldots, X_n$ with $X_i \sim F$ for $i = 1, \ldots, n$

Testing problem:

$$H_0 : F = F_0 \qquad \text{against} \qquad H_1 : F \neq F_0,$$

where F_0 is a given and known distribution, e.g. $N(\mu_0, \sigma_0^2)$.

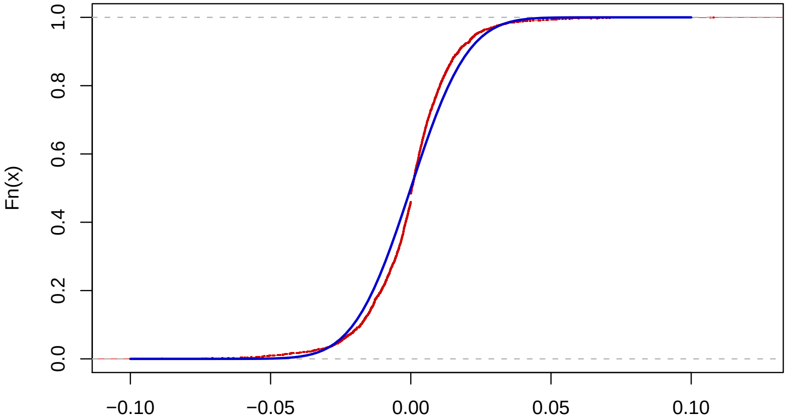
Idea of the test: Comparison of the empirical distribution function with F_0.

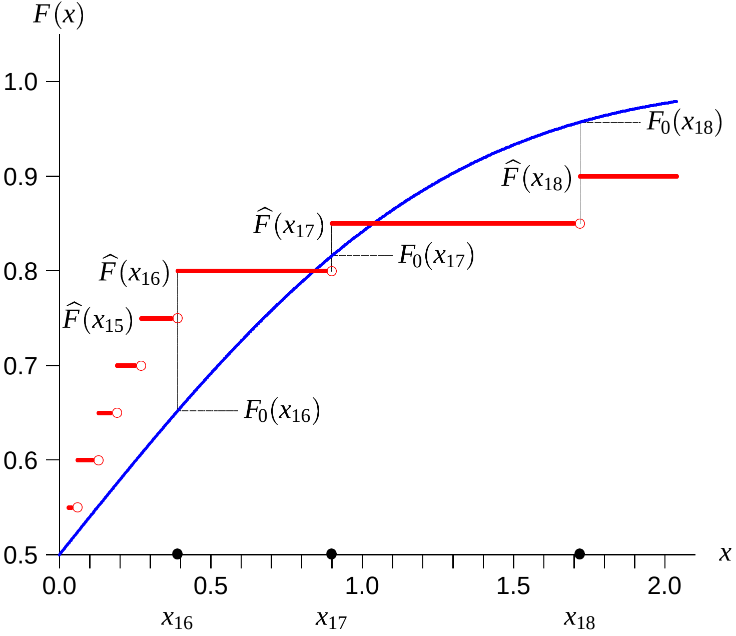
Distribution function:

$$F_0(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

- $0 \leq F_0(x) \leq 1$;
- $F_0(x)$ is a non-decreasing function;
- $F_0(x)$ is right-continuous.

Example: Distribution function of the normal distribution (blue) and the empirical distribution function (red) for DAX returns




$F(x)$
1.0
0.9
0.8
0.7
0.6
0.5
$F_0(x_{18})$
$\widehat{F}(x_{18})$
$\widehat{F}(x_{17})$
$F_0(x_{17})$
$\widehat{F}(x_{16})$
$\widehat{F}(x_{15})$
$F_0(x_{16})$
x
0.0
0.5
1.0
1.5
2.0
x_{16}
x_{17}
x_{18}

Test statistic: $D = \max_{x \in \mathbb{R}} |\hat{F}(x) - F_0(x)|$

The distribution of D under H_0 is a non-standard distribution and is independent from F_0 if F_0 is continuous!
Decision: using the *p-value*-approach.

In practice: Let F_0 be continuous and $x_1 \leq x_2 \leq \ldots \leq x_T$.

$$\rightsquigarrow \quad D = \max_{1 \leq t \leq T} \left\{ \hat{F}(x_t) - F_0(x_t), F_0(x_t) - \hat{F}(x_{t-1}) \right\},$$

where $\hat{F}(x_0) := 0$.

Now: F_0 is a non-predetermined distribution, but a class of distributions, e.g. $N(\cdot,\cdot)$.

Testing problem:

$$H_0 : F \in \mathcal{F}_0 := \left\{ F_0\left(\frac{x-\mu}{\sigma}\right) : \mu \in \mathbb{R}, \sigma > 0 \right\}$$

$$H_1 : F \in \mathcal{D} - \mathcal{F}_0\,,$$

where F_0 is known (e. g. $F_0 = \Phi$).

Modified Kolmogorov-Smirnov Test

Test statistic: $D^* = \max_{x\in\mathbb{R}} \left| \hat{F}(x) - \Phi\big((x-\hat{\mu})/\hat{\sigma}\big) \right|$

If $D^* > c^*$, then H_0 is rejected.

Example:

For the DAX-Index with $F_0 = \mathcal{N}(5.7854 \cdot 10^{-6}; 2.5551 \cdot 10^{-4})$ we get

```
> ks.test(rdax, "pnorm", mean=mean(rdax), sd= sd(rdax))

	One-sample Kolmogorov-Smirnov test

data:  rdax
D = 0.0736, p-value = 1.067e-12
alternative hypothesis: two-sided
```

$\Rightarrow$ The returns are not normally distributed.

Power of a test

- Parametric test:

$$H_0: \ \vartheta \in \Theta_0 \quad \text{vs} \quad H_1: \ \vartheta \in \Theta_1$$

  with $\Theta_0 \cup \Theta_1 = \Theta \subseteq \mathbb{R}$.
- Performance measures of a test:
  a) Prob. of type I error should not exceed α.
  b) Prob. of type II error should be a small as possible.

Prob. of rejection H_0 depending on the true value of the parameter

$$G(\mu) = P(V \in \text{rejection area} H_0 | \mu) = P(\{H_1\}|\mu)$$

- Is $\vartheta \in \Theta_0$ so we made a wrong decision $(\{H_1\}|H_0)$.
- The power function is in this case the prob. of type I error:

$$G(\mu) = P(\{H_1\}|\mu) \leq \alpha \text{ for all } \mu \in \Theta_0$$

- Is $\vartheta \in \Theta_1$ so we made the correct decision $(\{H_1\}|H_1)$.
- The prob. of type II error :

$$G(\mu) = P(\{H_1\}|\mu) \leq 1 - \beta \text{ for all } \mu \in \Theta_1$$

Power function of the test for the mean

Assumption: α and n are fixed, normal distribution and σ^2 is known.

$$
\begin{aligned}
G(\mu) &= P(V \in \text{ rejection area } H_0|\mu) \\
&= P(\{H_1\}|\mu) \\
&= 1 - P(V \in \text{ non-rejection area } H_0|\mu) \\
&= 1 - P(\{H_0\}|\mu)
\end{aligned}
$$

two-sided test

$$H_0 \;:\; \mu = \mu_0 \text{ vs. } H_1 \;:\; \mu \neq \mu_0$$

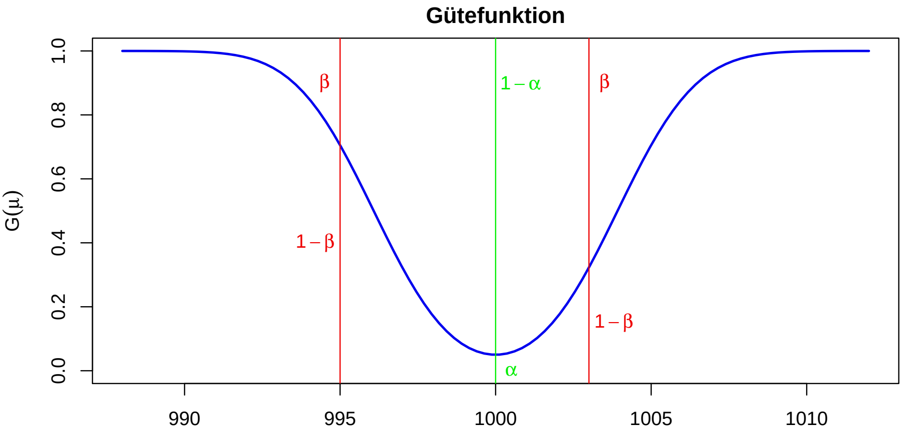
H_0 is correct only if $\mu = \mu_0$.

$$\begin{aligned}
G(\mu) =& 1 - P(-z_{1-\alpha/2} \leq V \leq z_{1-\alpha/2} \mid \mu) \\
=& 1 - P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \leq z_{1-\alpha/2} \mid \mu\right) \\
=& 1 - P\left(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq (\bar{X} - \mu_0) \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \mid \mu\right) \\
=& 1 - P\left(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} + \mu_0 - \mu \leq (\bar{X} - \mu) \leq z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} + \mu_0 - \mu \mid \mu\right) \\
=& 1 - P\left(-z_{1-\alpha/2} + \frac{\mu_0 - \mu}{\sigma}\sqrt{n} \leq \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \leq z_{1-\alpha/2} + \frac{\mu_0 - \mu}{\sigma}\sqrt{n} \mid \mu\right)
\end{aligned}$$

Since μ is the true mean, it holds $\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$.

$$
\begin{aligned}
G(\mu) &= 1 - \Big[P\Big(\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \leq z_{1-\alpha/2} + \frac{\mu_0-\mu}{\sigma}\sqrt{n}\Big) \\
&\quad -\Big[P\Big(-z_{1-\alpha/2} + \frac{\mu_0-\mu}{\sigma}\sqrt{n} \leq \frac{\bar{X}-\mu}{\sigma}\sqrt{n}\Big)\Big] \\
&= 1 - \Big[\Phi\Big(z_{1-\alpha/2} + \frac{\mu_0-\mu}{\sigma}\sqrt{n}\Big) - \Phi\Big(-z_{1-\alpha/2} + \frac{\mu_0-\mu}{\sigma}\sqrt{n}\Big)\Big]
\end{aligned}
$$

$$
G(\mu) = \begin{cases} \alpha = P(\{H_1\}|\mu), & \text{for } \mu = \mu_0 \\ 1-\beta(\mu) = P(\{H_1\}|\mu), & \text{for } \mu \neq \mu_0 \end{cases}
$$


Gütefunktion
G(μ)
0.0
0.2
0.4
0.6
0.8
1.0
990
995
1000
1005
1010
β
1 − α
β
1 − β
1 − β
α

Example: target filling capacity $\mu_0 = 1000$. Let $\sigma = 10$, $\alpha = 0.05$, $n = 25$. What is the prob. of type II error, if the true filling capacity is $\mu = 1002$?

$$
\begin{aligned}
G(1002) &= 1 - \Big[\Phi\Big(1.96 + \frac{1000-1002}{10}\sqrt{25}\Big) \\
&\quad -\Phi\Big(-1.96 + \frac{1000-1002}{10}\sqrt{25}\Big)\Big] = 0.170066 = 1-\beta
\end{aligned}
$$

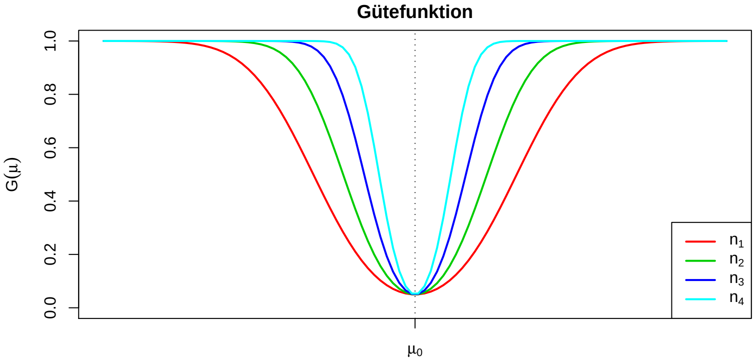
$$P(\{H_0\}|\mu = 1002) = \beta = 1 - G(1002) = 0.83$$

We will not detect the deviation of 2ml from the target capacity of 1000ml in 83% of the cases!!!

Let $\mu = 989$. Then $G(989) = 0.9998$ and $\beta = 0.0002$.

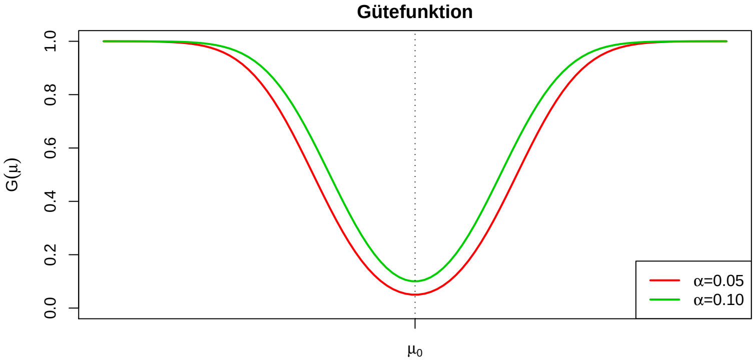
If the true capacity is $\mu = 989$, then we will NOT detect it only in 0.02% of all samples of size $n = 25$

Power function for different sample sizes $n_1 < n_2 < n_3 < n_4$



$\rightsquigarrow$ increasing the sample size reduces the prob. of type II error (ceteris paribus) .

Power function as a function of α



$\rightsquigarrow$ increasing the prob. of type I error reduces the prob. of type II error (ceteris paribus)

$\rightsquigarrow$ both probabilities cannot be reduced simultaneously!!!