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 $F(x)$.

distribution, expectation, variance ? But instead

make statements using a sample.

Universität Augsburg

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:

  $\{64; 69\}$.

  - 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).
  - the parameter (prob. of functioning)
  - 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)

is a scalar number

Aim: Using the sample data we determine an estimator (value) for the unknown parameter ϑ $\Longrightarrow \hat{\vartheta}$

p for party A?

Example:

- estimate the fraction of the voters, who voted for a particular party, using an exit-poll sample $\Rightarrow \hat{p}$ – estimator of the true unknown fraction
- 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%. $\Rightarrow p \in (71\%; 77\%)$ with prob 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 $\hat{\vartheta}$.

$\vartheta = p$. g - ? $g(A, A, B, A, \ldots) \to \hat{p}$

Universität Augsburg

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 $E(X)$ X - body height ; E(X) - expected body height.

Estimator: $\hat{\vartheta} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ sample: $x_1 \ldots x_n$ (1,70; 1,76 ...)

Estimate: sample mean $\hat{\vartheta} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ $\widehat{E(X)} = \hat{\mu}$ = average.

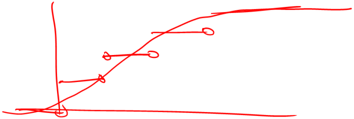
- Variance
  - $Var(X) = \sum_i (x_i - E(X))^2 P(X = x_i)$ or
    $Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$

$\hat{p}$, $\hat{\lambda}$ - how to develop a general approach to estimate a parameter?

Estimator for $Var(X)$

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

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



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.

bulbs: 20 bulbs. 1, 1, 0, 1, 1. 0 . . 1

depends on p

$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 1 \ldots X_{20} = 1) \Longrightarrow$ w.r.t. p.

Universität Augsburg

Typical ML-estimation

1. Write down the likelihood function : $L_\vartheta(x_1, \ldots, x_n)$
2. Take logs (optional): $\ln L_\vartheta(x_1, \ldots, x_n)$ $\Longrightarrow$ to get rid of $\prod_{i=1}^{n}$ (difficult to take derivatives).
3. Take derivatives: $\frac{\partial}{\partial\vartheta}[\ln]L_\vartheta(x_1, \ldots, x_n) \stackrel{!}{=} 0$
4. (usually thete are many parameters!)

Binomial

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 $\underbrace{X_1, ..., X_n}$ are independent and $X_i \sim B(1,p)$ for $i = 1, .., n$.

before treatment.

$\{0,1\}$ after treatment.

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

$$L_p(x_1, \dots, x_n) = \prod_{i=1}^{n} f_i(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \quad \rightarrow \text{max w.r.t } p$$

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

Taking logarithm yields

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

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 $\hat{p} = \bar{x}$. $\Longrightarrow$ 100 patients ; 80 success $\Rightarrow \hat{p} = \frac{80}{100} = 0{,}8$

sample $= \frac{1}{100}(1 + 1 + 0 + 1 \ldots)$

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 $\hat{\Theta} = \bar{X}$.

lifetime.

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

$$L_\lambda(x_1, \ldots, x_n) = \prod_{i=1}^{n} \overbrace{\lambda \exp(-\lambda\, x_i)}^{\text{density function for exponential.}} = \lambda^n \exp(-\lambda\, n\, \bar{x})\,,$$

$$\ln L_\lambda(x_1, \ldots, x_n) = n \ln\lambda - n\,\lambda\,\bar{x}.$$

The solution is

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

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

bulbs with life times: 7, 5, 8 . . . 15 years.

$$\hat{\lambda} = \frac{1}{\frac{1}{10}(7+5+8+\ldots 15)} \Rightarrow P(X > 15 \text{ years})$$

Universität Augsburg

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)

$$Bias(\hat{\vartheta}) = E(\hat{\vartheta}) - \vartheta \quad = 0$$

using many samples ⇒ we will estimate the true parameter without an error.

- Variance: should be small $\rightsquigarrow$ efficiency

$\parallel$ Θ of unbiased

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

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

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

measure of shift. measure of precision.

Universität Augsburg

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 $E(X_i) = \mu$?

$E\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n}\left(E(X_1) + \ldots + E(X_n)\right)$
$= \frac{1}{n}(\mu + \ldots + \mu) = \mu.$

- $E(\hat{\Theta}) = E(\bar{X}_n) = \mu$
  $\Rightarrow$ 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
- $E(\hat{\Theta}'') = E\left(\frac{1}{n-1}\sum\limits_{i=1}^{n} X_i\right) = \frac{1}{n-1}\sum\limits_{i=1}^{n} E(X_i) = \frac{1}{n-1}\sum\limits_{i=1}^{n} \mu =$
  $\frac{n}{n-1} \cdot \mu \underset{n\to\infty}{\longrightarrow} \mu$
  $\Rightarrow$ asymptotically unbiased.

Which estimator better?

Universität Augsburg

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:

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

$\Rightarrow$ $\hat{\Theta}$ is on average more precise than $\hat{\Theta}'$

Example:

$\Rightarrow$ $Var\left(\frac{1}{n}(X_1 + \dots + X_n)\right) = \frac{1}{n^2}(Var X_1 + \dots + Var X_n) = \frac{n\sigma^2}{n^2} = \sigma^2/n$

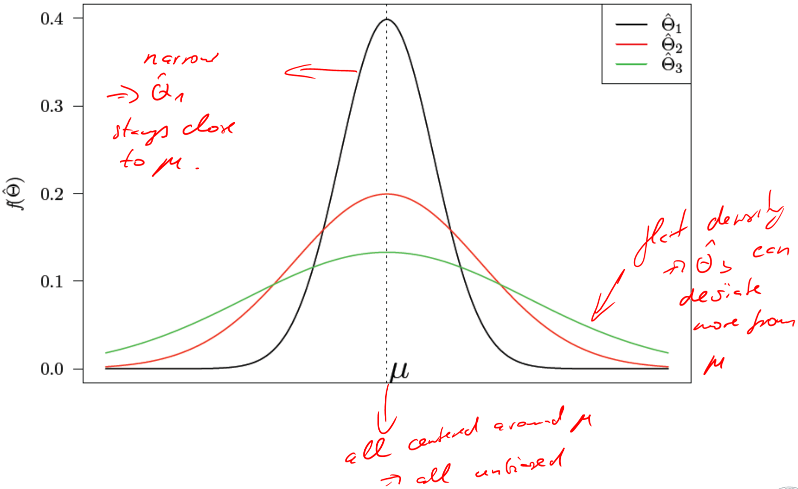
$$Var(\hat{\Theta}) = Var(\bar{X}) = \frac{\sigma^2}{n} <$$

$$< Var(\hat{\Theta}') = Var\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}$$

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

$\Rightarrow$ $\bar{X}$ is more "precise" than $\frac{X_1 + X_n}{2}$, because of more information used for estimation

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



Universität Augsburg

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

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

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

Probability that we miss μ by more than C.

$$P(\underbrace{|\hat{\Theta}_n - \theta|}_{\text{absolute distance between what we measured for n people and expected body height.}} \geq c) \underset{n\to\infty}{\longrightarrow} 0$$

Asymptotically if sample size increases to infinity!

absolute distance between what we measured for n people and expected body height.

difference is larger than c.

Universität Augsburg

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

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

$\Rightarrow E(\bar{X}_n) = \mu$ because unbiased.

The weak low of large numbers

$\bar{X}_n \Rightarrow Var(\bar{X}_n) = \frac{\sigma^2}{n}$.

1. Tschebyscheff inequality

$\bar{X}_n$ $\bar{X}_n$

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

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

$$P(|\bar{X}_n - \mu| \geqq c) \leq \frac{\sigma^2}{n \cdot c^2} \rightarrow 0, \text{ as } n \rightarrow \infty.$$

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

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

$\Rightarrow$ increasing the sample size makes the mean to converge to the expectation.

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

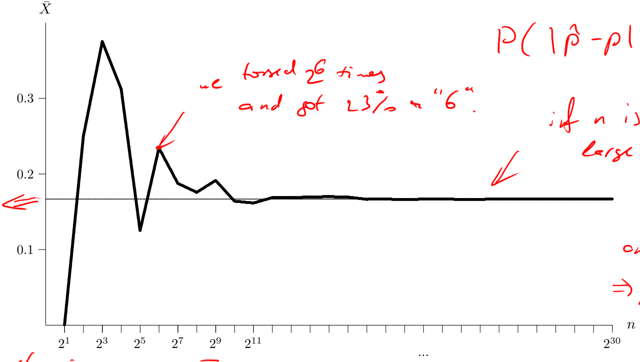
$$X_i = \begin{cases} 1 & \text{if you toss a 6} \\ 0 & \text{if you toss other number} \end{cases}$$

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

we know from ML, $\hat{p} = \bar{X} \Rightarrow$ fraction of "6" by n tosses.

We know that the true value is $1/6 = p$.

$P(|\hat{p} - p| > c) \to 0, \; n \to \infty$.



we tossed 2^6 times and got 23% in "6".

$p = 1/6$

if n is large, then large deviations between $\hat{p}$ and $p = 1/6$ are very improbable $\Rightarrow \hat{p}$ is consistent for p

Here: $p = 1/6$ is known. For more complex models θ is unknown.
Practice: if you increase sample, then the estimator will closer and closer to the true value.

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.

$\Rightarrow$ ML $\Rightarrow$ need a distribution (say body height is normal)
$\Rightarrow$ what happens if the assumption is not correct?

μ - expected body weight. $\Rightarrow$ point estimator is $\bar{X}$.
$\Rightarrow$ interval estimation $\mu \in (150; 190)$ with prob. 95%.
to derive this we need the distribution of $\bar{X}$.

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

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

$X_1 \sim N$, $X_2 \sim N$ .. $X_n \sim N$
$\Rightarrow \frac{1}{n}(X_1 + .. + X_n) \sim N$.

- 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

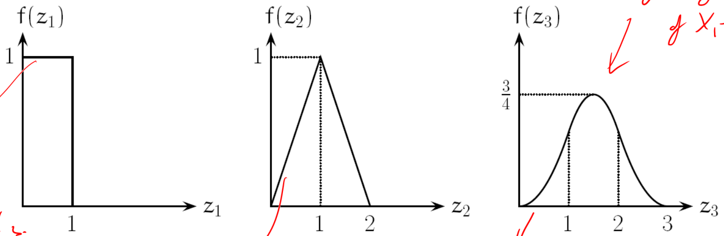
$X_1 \sim F_1$ $X_n \sim F_n$
$\bar{X} \sim ?$

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;
$Z_1 = X_1, \;\; Z_2 = X_1 + X_2, \;\; Z_3 = X_1 + X_2 + X_3$

density of $X_1 + X_2 + X_3$



densities of X_1, X_2, X_3 (clearly non-normal)

density of $X_1 + X_2$

already Bell-shaped despite of

Universität Augsburg

$X_1 \sim F_1, \quad Y_n \sim F_n$ with $E X_i = \mu$, $Var X_i = \sigma^2$

1. Standardisation:

$E(\bar{X}_n)$

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

$\sqrt{Var(\bar{X}_n)} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$.

2.

Central Limit Theorem

distribution of Y_n at x.

$$\lim_{n\to\infty} \overbrace{P(Y_n \leq x)}^{\text{Definition of the}} = \lim_{n\to\infty} P\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq x\right) = \Phi(x)$$

distribution of the standard normal distribution $\mathcal{N}(0,1)$

3. We write

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

d sample size goes to ∞.

μ - expected body height. X - body height.

$X = \underbrace{\frac{1}{n}}_{= \text{factors}} \sum_{i=1}^{n} \text{factors}_i =$

Universität Augsburg

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?

$X_i = \begin{cases} 1, & \text{if } i\text{th part is replaced} \\ 0, & \text{not} \end{cases}$

$$P\left(10 < \overbrace{\underbrace{\sum_{i=1}^{100} X_i}_{\text{number of parts to be replaced.}}}^{B(100;\frac{1}{6}).} \leq 20\right) = F_{B(100,\frac{1}{6})}(20) - F_{B(100,\frac{1}{6})}(10)$$

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

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

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

$$P\left(\frac{10}{100} < \frac{\sum_{i=1}^{100} X_i}{100} \leq \frac{20}{100}\right) = P\left(0.1 < \bar{X} \leq 0.2\right)$$

should be approximately normal according to CLT.

$$= P\left(\underbrace{\sqrt{100}\,\frac{0.1-\frac{1}{6}}{\sqrt{\frac{5}{36}}}}_{=-1.7888} < \underset{=\sqrt{n}}{\sqrt{100}}\,\frac{\bar{X}-\overset{E(X_i)}{\frac{1}{6}}}{\underset{\sigma}{\sqrt{\frac{5}{36}}}} \leq \underbrace{\sqrt{100}\,\frac{0.2-\frac{1}{6}}{\sqrt{\frac{5}{36}}}}_{=0.8944}\right)$$

$$\overset{\text{CLT!}}{\approx} \Phi(0.8944) - \Phi(-1.7888)$$

$$= \Phi(0.8944) - 1 + \Phi(1.7888)$$

$$= 0.8143 - 1 + 0.9631$$

$$= 0.7774$$

cumulative distrib function of $N(0,1)$.

$\Rightarrow$ with $\approx 78\%$ one has to replace between 10 and 20 parts

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))

X-body height ~ N   Y - body weight N.
BMI = body mass index = $\frac{\text{weight}}{\text{height}^2}$.

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* $\sqrt{n}(\boldsymbol{z}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ *and if* $\boldsymbol{c}(\cdot) : \mathbb{R}^k \to \mathbb{R}^J$ *is a continuous function not involving* n*, then*

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

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 $\alpha \in (0, 1)$.

boundaries are computed using sample data.

The interval $[T_1(\boldsymbol{X}), T_2(\boldsymbol{X})]$ with

unknown parameter.

$$P_\vartheta(T_1 \leq \vartheta \leq T_2) = \underbrace{1 - \alpha}_{95\%,\ 90\%,\ 95\%} \quad \forall \vartheta \in \Theta \qquad (*)$$

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

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

and $\big(-\infty, T_2(\boldsymbol{X})\big]$ is a one-sided upper confidence interval

at the confidence level $1-\alpha$, if

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

risk → lower bounding is irrelevant → interest in the highest possible risk!

$P(\text{risk} \leq T_2) = 0{,}95.$

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 $\bar{X}$ = natural starting point (point estimator)!

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

build the CI around the point estimator $\bar{X}$.

$$[\overbrace{\bar{X} + c_1 \frac{\sigma}{\sqrt{n}}}^{T_1}, \overbrace{\bar{X} + c_2 \frac{\sigma}{\sqrt{n}}}^{T_2}]$$

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 $c_2 = c = -c_1$.

$$P\{\bar{X} - c \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + c \frac{\sigma}{\sqrt{n}}\} = 1-\alpha.$$

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

standardization $\dfrac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} \sim N(0,1)$

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

definition of the distrib. function at c

distribution at $-c$.

according to the definition of a CI.

This implies $\boxed{c = z_{1-\alpha/2}}$ → quantile of $N(0,1)$ at $1-\alpha/2$.

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

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

$X_1 = 110; \; X_2 = 80 \; - \; - \qquad X_{100} = 75 \Rightarrow \bar{X} = 100.$

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 $[80.4, 119.6]$. ⇒ expected demand lies with prob of 95% ⟶ ?.

⇒ store 120 items ⇒ the demand will be satisfied with 97,5% probability. ⇒ 2,5% prob that he has more orders.

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

true value

estimator

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

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 $1-\alpha$

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

quantiles of the t_{n-1} distribution.

very close to $z_{1-\alpha/2} = 1{,}65$

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

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

R : qnorm(0,95) $\Rightarrow$ 1,6

iii) Confidence intervals for σ^2

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

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.

So far: $-c_1 = c = c_2$ w/N

$\to c_1 \leftrightarrow c_2$

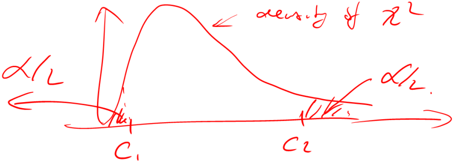
Since the χ^2–distribution is not symmetric, the CI for σ^2 is also not symmetric. We take

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

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 $c_1 = \chi^2_{n-1;\alpha/2}$.



both limits are quantiles of the χ^2 distribution

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

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

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

up to now: we needed a distributional assumption

now: Asymptotic confidence intervals

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

is unknown, because the distribution is unknown.

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

the parameter lies between T_1 and T_2 with $1-\alpha$ only for infinite samples.

Example: Let $X_1, ..., X_n$ be iid with $E(X_i) = \mu$ for all $i \ge 1$. The objective is a CI for μ. (but not normal).

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

$X_1 \dots X_n$ not normal $\Rightarrow \bar{X}$ not normal.

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.

$\not\sim N(0,1)$, but $\to N(0,1)$ by CLT

Universität Augsburg

For this reason one applies the CLT.

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

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

$\Rightarrow \Phi(c)$ $\Rightarrow \Phi(-c)$.

CLT

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

as before.

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

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

the unknown μ lies in this interval with prob. of approx. $1-\alpha$.

Rule of thumb: $n \ge 30$