NMAI059 Probability and statistics 1
Class 12

Robert Šámal

Overview

Statistics – point estimation

Statistics – interval estimation

Hypothesis testing

Sample mean & variance

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

$$\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \longrightarrow \text{unbiased}$$

lower MSE

Maximal likelihood method, ML)

p / f

unknown parameter

known data

- **Maximal likelihood method:**
  $\boxed{\text{choose } \vartheta \text{ that maximizes } L(x;\vartheta).}$ → $\hat{\vartheta}$ — usually done by differs.
- for convenience we put $\ell(x;\vartheta) = \log(L(x;\vartheta))$
- by independence of X_1, X_2, etc. we have

$$L(x;\vartheta) = L(x_1;\vartheta)\dots L(x_n;\vartheta)$$
$$\ell(x;\vartheta) = \ell(x_1;\vartheta) + \dots + \ell(x_n;\vartheta)$$

do not know

$\hat{p} = \frac{9}{20}$

P

Bin(20,p)	0.2	0.3	0.4	0.45	0.5	0.55	0.6
7	0.0545	0.1643	0.1659	0.1221	0.0739	0.0366	0.0146
8	0.0222	0.1144	0.1797	0.1623	0.1201	0.0727	0.0355
9	0.0074	0.0654	0.1597	0.1771	0.1602	0.1185	0.071
10	0.002	0.0308	0.1171	0.1593	0.1762	0.1593	0.1171
11	0.0005	0.012	0.071	0.1185	0.1602	0.1771	0.1597
12	0.0001	0.0039	0.0355	0.0727	0.1201	0.1623	0.1797
13	0	0.001	0.0146	0.0366	0.0739	0.1221	0.1659
14	0	0.0002	0.0049	0.015	0.037	0.0746	0.1244

know → 9

depends on p

20 measurements, indep. 0/1, p – prob of 1

find max.

9 said yes

pmf

$$\boxed{P_p(X=k) = \binom{20}{k} p^k (1-p)^{20-k}}$$

ML – further illustration $N(\mu, \sigma^2)$ $\vartheta = (\mu, \sigma)$

$x = (x_1, \dots, x_n)$ numbers — realizations of $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\left(\frac{x_i - \mu}{\sigma}\right)^2 / 2}$ ..... formula for pdf of

$L(x_i; \vartheta) \Rightarrow \ell(x_i; \vartheta) = -\left(\frac{x_i - \mu}{\sigma}\right)^2 / 2 - \log \sigma - \log \sqrt{2\pi}$

$\ell(x; \vartheta) = -\sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2 / 2 - n \log \sigma - n \log \sqrt{2\pi}$

$\boxed{\text{FIND } \vartheta = (\mu, \sigma) \text{ THAT MAXIMIZES THIS}}$

$\boxed{\text{TO DO THIS DIFFERENT.}}$

$\frac{\partial \ell}{\partial \mu} = + \sum_{i=1}^{n} \cancel{2} \left(\frac{x_i - \mu}{\sigma}\right) \cdot \frac{1}{\cancel{2}} \cdot \frac{1}{\sigma} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$ EQ1 $= 0 \Rightarrow \mu = \frac{x_1 + \dots + x_n}{n} = \bar{x}_n$

$\boxed{\hat{\mu} := \bar{x}_n}$

$\frac{\partial \ell}{\partial \sigma} = + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} \frac{\cancel{2}}{\cancel{2}} - \frac{n}{\sigma} = 0$ EQ2 $\Rightarrow \sigma^2 = \frac{1}{n} \sum (x_i - \mu)^2$

$\boxed{\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x}_n)^2}$

Overview

Statistics – point estimation

Statistics – interval estimation

Hypothesis testing

Interval estimation

unknown parameter

- Instead of estimating by one number we compute from our data an interval $[\hat{\Theta}^-, \hat{\Theta}^+]$

Definition

Let $\hat{\Theta}^-$, $\hat{\Theta}^+$ be random variables that depend on the random sample $X = (X_1, \ldots, X_n)$ from distribution F_ϑ. These random variables describe a $1-\alpha$ confidence interval, if

$$P(\underbrace{\hat{\Theta}^-}_{A} \leq \vartheta \leq \underbrace{\hat{\Theta}^+}_{B}) \geq 1-\alpha.$$

NOT A PROB. STATEMENT ABOUT ϑ !

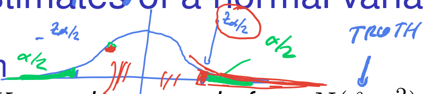
- these are two-sided estimates
- one-sided: $[\hat{\Theta}^-, \infty)$ or $(-\infty, \hat{\Theta}^-]$

ϑ IS A FIXED PARAM. THAT WE DON'T KNOW

Interval estimates of a normal variable

pdf of $N(0,1)$

we measure temper.
const. of thermometer → σ
read temp. ϑ



Theorem

$X_1, \dots, X_n$ *random sample from* $N(\vartheta, \sigma^2)$.

σ ***is known***, *we need to estimate* ϑ, *we choose* $\alpha \in (0,1)$.

Let $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. *We put* $\hat{\Theta}_n := \bar{X}_n$ *and*

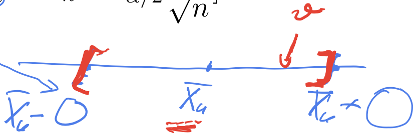
$z_{\alpha/2} := \Phi^{-1}(1-\frac{\alpha}{2})$

$\bar{X}_n \sim N(\vartheta, \frac{\sigma^2}{n})$

$Z = \frac{\bar{X}_n - \vartheta}{\sigma/\sqrt{n}} \sim N(0,1)$

$$C_n := [\hat{\Theta}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \hat{\Theta}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Then $P(C_n \ni \vartheta) = 1 - \alpha$.



Důkaz.

$P(C_n \ni \vartheta) = P\left(|\bar{X}_n - \vartheta| \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$

$= P\left(\left|\frac{\bar{X}_n - \vartheta}{\sigma/\sqrt{n}}\right| \le z_{\alpha/2}\right) = P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2})$

Z :-)

$= (1 - \frac{\alpha}{2}) - \frac{\alpha}{2} = 1 - \alpha$

Interval estimates using CLT

not necessary $N(\vartheta, \sigma^2)$

Theorem

$X_1, \ldots, X_n$ *random sample from a distribution with mean* ϑ *and variance* σ^2.
$\boldsymbol{\sigma}$ ***is known*** *we need to estimate* ϑ, *we choose* $\alpha \in (0,1)$.
Let $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. *We put* $\hat{\Theta}_n := \bar{X}_n$ *and*

not necess. normal

$$C_n := [\hat{\Theta}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \hat{\Theta}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

$$Z_n := \frac{\bar{X}_n - \vartheta}{\sigma/\sqrt{n}}$$

Then $\boxed{\lim_{n\to\infty}} P(C_n \ni \vartheta) = 1 - \alpha.$

BUT CLT: $Z_n \xrightarrow{d} N(0,1)$

Proof $P(C_n \ni \vartheta) = P\left(|\bar{X}_n - \vartheta| \le z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$

$= P\left(|Z_n| \le z_{\alpha/2} \right) = F_{Z_n}(z_{\alpha/2}) - F_{Z_n}(-z_{\alpha/2}) = ? \xrightarrow{n\to\infty} \phi(z_{\alpha/2}) - \phi(-z_{\alpha/2}) = 1-\alpha$

Student t-distribution

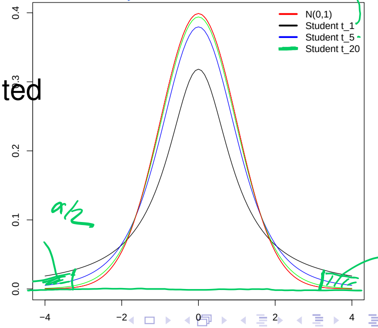
sample std. dev.

$\hat{S}_n = \sqrt{\hat{S}_n^2} = \sqrt{\frac{1}{n-1}\sum(X_i-\bar{X}_n)^2}$

- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ ... sample mean
- $\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ ... sample variance

does not depend on μ & σ

- Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.
- Then we know that $\boxed{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}} \sim N(0,1)$
- *Student t-distribution with $n-1$ degrees of freedom* is the distribution of r.v. $\boxed{\frac{\bar{X}_n - \mu}{\widehat{S}_n/\sqrt{n}}}$ $\xrightarrow{d} N(0,1)$
- Its cdf will be denoted Ψ_{n-1}
  It is tabulated, and implemented by computer sofware,
  in R: **pt**(x,n−1)



Int. estimates of normal variable using Student t

Theorem

$X_1, \dots, X_n$ random samples from $N(\vartheta, \sigma^2)$.
*σ **is not known**, we need to estimate ϑ, we choose $\alpha \in (0,1)$.*
Let $\boxed{\Psi_{n-1}(z_{\alpha/2}) = 1 - \alpha/2.}$ We put $\hat\Theta_n = \bar X_n$,
$\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar X_n)^2$ and

$$C_n := [\hat\Theta_n - z_{\alpha/2}\frac{\widehat{S}_n}{\sqrt n}, \quad \hat\Theta_n + z_{\alpha/2}\frac{\widehat{S}_n}{\sqrt n}]$$

Then $P(C_n \ni \vartheta) = 1 - \alpha$.

$\hat S_n$ instead of σ

$z_{\alpha/2}$ determined using Ψ_{n-1} inst. of Φ

$\bar X_n - \bigcirc \qquad \bar X_n \qquad \bar X_n + \bigcirc$

Proof

$P\left(|\bar X_n - \vartheta| \le z_{\alpha/2}\frac{\hat S_n}{\sqrt n}\right)$

$Z \sim \Psi_{n-1}$

$P\left(\left|\underbrace{\frac{\bar X_n - \vartheta}{\hat S_n/\sqrt n}}_{=:Z}\right| \le z_{\alpha/2}\right) = \Psi_{n-1}(z_{\alpha/2}) - \Psi_{n-1}(-z_{\alpha/2}) = \dots = 1-\alpha$

Overview

Statistics – point estimation

Statistics – interval estimation

Hypothesis testing

Intro to Hypothesis testing

- Is our coin fair? H_0 : yes
- Is our die fair? H_0 : yes
- Is the modified code faster then original? H_0 : no
- Is the medical treatment X good? (Better than placebo, better than Y, . . .) H_0 : no
- Are left-handed people better at boxing? H_0 : no

- two hypothesis: H_0, H_1
- H_0 – *null hypothesis* – default, conservative model, “unsurprising”
- H_1 – *alternative hypothesis* – alternative model “remarkable fact”, if true

Hypothesis testing – illustration

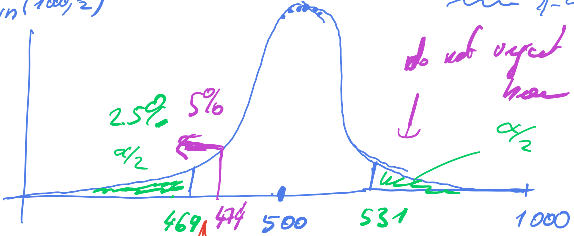
$n = 1000$

- We want to test, if a coin is fair.
- We toss it n-times, we get head S-times.
- If $|S - n/2|$ is too large, we declare the coin not to be fair.

$P(S=0) = 2^{-n}$ very small

pmf $Bin(1000, \frac{1}{2})$

$1-\alpha$ interval est.



$\alpha = 5\%$

472

say: unfair if $S < \frac{n}{2} - c$

Remark Suppose, we prefer heads.

Hypothesis testing – illustration

- We want to test, if a coin is fair.
- H_0: it is fair
- H_1: not fair ("Scientists discovered, that casino XY uses loaded coin.")
- Results: Reject H_0/don't reject H_0
- Type I error: false rejection. We reject H_0, even if it is true. Embarassing.
- Type II error: false non-rejection. We don't reject H_0, even if it is false. Unused opportunity.
- Need to find k such that we will reject H_0 if $|S - n/2| > k$.

Hypothesis testing – general approach

- We choose an appropriate statistical model.
- We choose *significance level* α: prob. of false rejection of H_0. Typically $\alpha = 0.05$ (medicine/psychology – much less in high-energhy physics).
- We determine *test statistics* $S = h(X_1, \dots, X_n)$, that we will determine from the measured data.
- We determine *rejection region* – set W.
- We measure $x_1, \dots, x_n$ – so-called realizations of $X_1, \dots, X_n$.
- Decision rule: we reject H_0 iff $h(x_1, \dots, x_n) \in W$.
- $\alpha = P(h(X) \in W; H_0)$
- $\beta = P(h(X) \notin W; H_1)$ ... *strength of the test*

- often we do not choose α in advance but compute so-called p-*value:* minimal α, for which we would reject H_0.

Hypothesis testing – an example

- $X_1, \ldots, X_n$ random sample from $N(\vartheta, \sigma^2)$
- σ^2 known
- $H_0 : \vartheta = 0 \qquad H_1 : \vartheta \neq 0$

Hypothesis testing – an example

- $X_1, \ldots, X_{n_1}$ random sample from $Ber(\vartheta_X)$
- $Y_1, \ldots, Y_{n_2}$ random sample from $Ber(\vartheta_Y)$
- $H_0 : \vartheta_X = \vartheta_Y \qquad H_1 : \vartheta_X \neq \vartheta_Y$

p-hacking

- we first gain data, then look for interesting stuff
- – given enough data, there will be random coincidences
- even worse, we may test, until we get the desired outcome
- *reproducibility* – after exploratory analysis of the data we make an independent measurement and a confirmatory analysis.
- or we split the data in advance to a part for hypothesis formation and part for verification . . . simple example of cross validation