Statistics

L4: Estimation

Katarzyna Filipiak

Institute of Mathematics
Poznań University of Technology

2025/2026

Random sample

Aim of statistics: to draw conclusions about population from a set of observed data.

Sample – a known and measurable entity representing a population having unknown features

Observations: $x_1,\ x_2, \ldots,\ x_n$
(values of random variables $X_1, X_2, \ldots, X_n$)

Estimation

θ - unknown parameter in a population distribution, $f(x)$

Estimator of θ – statistic (the function) that describes the method of computing of the estimate of θ:

$$\widehat{\Theta} = \widehat{\Theta}(X_1, X_2, \dots X_n)$$

Estimator = statistic = random variable having its distribution!

Observations - the results of experiment: $x_1, x_2, \dots, x_n$

Estimate of θ:

$$\widehat{\theta} = \widehat{\Theta}(x_1, x_2, \dots x_n)$$

Point estimation – population mean

Observed feature in a population – random variable X with distribution with unknown population mean μ

Parameter (unknown): population mean μ

Sample: $X_1, X_2, \ldots, X_n$

Estimator of μ: sample mean, $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

Observations: $x_1, x_2, \ldots, x_n$

Estimate of μ: $\overline{x}$

$$\boxed{\widehat{\mu} = \overline{x}}$$

Example 1

The mechanical engineer who designed the physical-therapy device collected the following data for the amount of time (in hours) spent by the test patients using his machine:

$$8; \ 12; \ 26; \ 10; \ 23; \ 21; \ 16; \ 22; \ 18; \ 17; \ 36; \ 9.$$

Estimate the mean time spent until recovery by all patients who use the same therapy.

Interval estimation

θ – unknown parameter

$\widehat{\Theta}$ – parameter estimator

Definition

If

$$P\left(L(\widehat{\Theta}) < \theta < U(\widehat{\Theta})\right) = 1 - \alpha$$

then

$$\left(L(\widehat{\Theta}); U(\widehat{\Theta})\right)$$

is called a $(1-\alpha) \cdot 100\%$ confidence interval, and the probability $(1-\alpha)$ is a confidence level.

Sample mean distribution – revisited

(1) $X_i \sim N(\mu, \sigma)$, μ, σ – known:

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \quad \Rightarrow \quad \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

(2) X_i with μ, σ – known, distribution of X_i is arbitrary, sample is large:

$$\overline{X} \underset{\text{app}}{\sim} N(\mu, \frac{\sigma}{\sqrt{n}}) \quad \Rightarrow \quad \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \underset{\text{app}}{\sim} N(0, 1)$$

(3) $X_i \sim N(\mu, \sigma)$, μ – known, σ – unknown: $\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

Confidence interval for μ (1)

Assumption: $X_i \sim N(\mu, \sigma)$, μ - unknown, σ - known

We are $100(1-\alpha)\%$ confident, that the interval

$$\left(\overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

covers the true unknown **population** mean μ.

$z_{1-\alpha/2}$ – quantile of $N(0,1)$: $\texttt{qnorm}(1-\alpha/2)$

Confidence interval for μ (2)

Assumptions: arbitrary population distribution with unknown μ,
large sample ($n \geq 30$)

We are $100(1-\alpha)\%$ confident, that the interval

$$\left(\overline{X} - z_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}},\ \overline{X} + z_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}}\right)$$

covers the true unknown **population** mean μ.

$z_{1-\alpha/2}$ – quantile of $N(0,1)$: `qnorm(1 - α/2)`

Confidence interval for μ (3)

Assumption: $X_i \sim N(\mu, \sigma)$, μ - unknown, σ - unknown

We are $100(1-\alpha)\%$ confident, that the interval

$$\left(\overline{X} - t_{n-1,1-\alpha/2} \cdot \frac{S}{\sqrt{n}},\ \overline{X} + t_{n-1,1-\alpha/2} \cdot \frac{S}{\sqrt{n}}\right)$$

covers the true unknown **population** mean μ.

$t_{n-1,1-\alpha/2}$ – quantile of t_{n-1}: $\texttt{qt}(1-\alpha/2,\ n-1)$

Example 1 - cont.

The mechanical engineer who designed the physical-therapy device collected the following data for the amount of time (in hours) spent by the test patients using his machine:

8; 12; 26; 10; 23; 21; 16; 22; 18; 17; 36; 9.

Assuming normality of the distribution of time, estimate with 95% of confidence the mean time spent until recovery by all patients who use the same therapy.

Example 1 - cont.

Confidence intervals for μ in R

(1) $X_i \sim N(\mu, \sigma)$, σ – known:

$$\texttt{z.test}(\text{data}, \texttt{sigma.x} = \sigma, \texttt{conf.level} = 1 - \alpha)$$

(2) distribution of X_i is arbitrary, sample is large:

σ – known

$$\texttt{zsum.test}(\texttt{mean}(\text{data}), \quad \sigma \quad, n, \texttt{conf.level} = 1 - \alpha)$$

σ – unknown

$$\texttt{zsum.test}(\texttt{mean}(\text{data}), \texttt{sd}(\text{data}), n, \texttt{conf.level} = 1 - \alpha)$$

(3) $X_i \sim N(\mu, \sigma)$, σ – unknown:

$$\texttt{t.test}(\text{data}, \texttt{conf.level} = 1 - \alpha)$$

CAUTION! For (1) and (2) - `BSDA` package required

Example 1 - cont.

The mechanical engineer who designed the physical-therapy device collected the following data for the amount of time (in hours) spent by the test patients using his machine:

8; 12; 26; 10; 23; 21; 16; 22; 18; 17; 36; 9.

Assuming normality of the distribution of time, estimate with 95% of confidence the mean time spent until recovery by all patients who use the same therapy.

Example 1 - cont.

Point estimation – population variance

Observed feature in a population – random variable X with distribution with unknown population variance σ^2

Parameter (unknown): population variance σ^2

Sample: $X_1, X_2, \ldots, X_n$

Estimator of σ^2: sample variance, $S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$

Observations: $x_1, x_2, \ldots, x_n$

Estimate of σ^2: s^2

$\boxed{\widehat{\sigma}^2 = s^2}$

Example 1 - cont.

The mechanical engineer who designed the physical-therapy device collected the following data for the amount of time (in hours) spent by the test patients using his machine:

8; 12; 26; 10; 23; 21; 16; 22; 18; 17; 36; 9.

Estimate the standard deviation of the time spent until recovery by all patients who use the same therapy.

Confidence interval for σ^2

Assumption: $X_i \sim N(\mu, \sigma)$, σ - unknown

We are $100(1-\alpha)\%$ confident, that the interval

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

covers the true unknown **population** variance σ^2.

$\chi^2_{n-1,\beta}$ – quantiles of χ^2_{n-1}: $\texttt{qchisq}(\beta, n-1)$

Confidence interval for σ^2 in R:

$$\texttt{sigma.test}(\text{data}, \texttt{conf.level} = 1-\alpha)$$

CAUTION! `TeachingDemos` package required

Example 1 - cont.

The mechanical engineer who designed the physical-therapy device collected the following data for the amount of time (in hours) spent by the test patients using his machine:

8; 12; 26; 10; 23; 21; 16; 22; 18; 17; 36; 9.

Assuming normality of the distribution of time, estimate with 95% of confidence the variance and standard deviation of time spent until recovery by all patients who use the same therapy.

Example 1 - cont.

Point estimation – population proportion

Observed feature in a population – random variable X with distribution $\mathrm{bin}(1, p)$ with unknown probability of success p

Parameter (unknown): population proportion p

Sample: $X_1, X_2, \dots, X_n$

Estimator of p: sample proportion, $\widehat{p} = \frac{T}{n}$

$T = \sum_{i=1}^{n} X_i$ - number of "successes" in a sample

Observations: $x_1, x_2, \dots, x_n$

Estimate of p: $\widehat{p}$

Example 2

A school district is trying to determine its students' reaction to a proposed dress code. To do so, the school selected a random sample of 150 students and questioned them. If 70 were in favor of the proposal, then estimate the proportion of all students who are in favor.

Confidence interval for p

Assumption: large sample ($n \geq 100$)

We are $100(1-\alpha)\%$ confident, that the interval

$$\left(\widehat{p} - z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)$$

covers the true unknown **population** proportion p.

$z_{1-\alpha/2}$ – quantile of $N(0,1)$: $\texttt{qnorm}(1-\alpha/2)$

Confidence interval for p in R:

$\texttt{binom.test}(T, n, \texttt{conf.level} = 1-\alpha)$ (exact)
$\texttt{prop.test}\ (T, n, \texttt{conf.level} = 1-\alpha)$ (approximate)

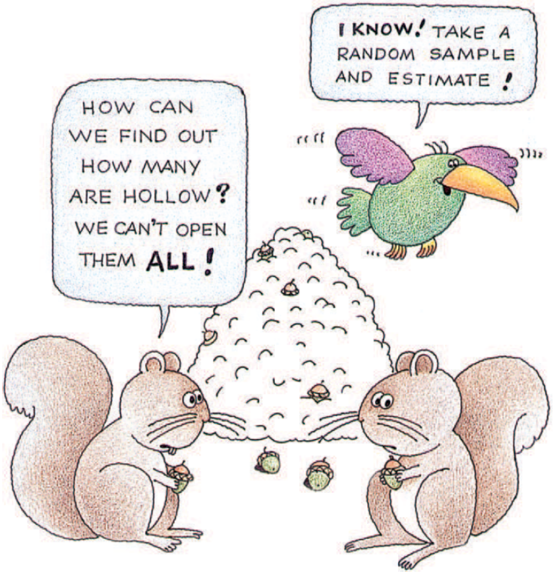
Example 2 - cont.

A school district is trying to determine its students' reaction to a proposed dress code. To do so, the school selected a random sample of 150 students and questioned them. If 70 were in favor of the proposal, then estimate with 99% of confidence the true proportion of all students who are in favor of the proposal.

Example 2 - cont.

Example 3

On December 24, 1991, *The New York Times* reported that a poll indicated that 46% of the population was in favor of the way that President Bush was handling the economy, with a margin of error of $\pm 3\%$. What does this mean? Can we conclude how many people were questioned if it is known, that the standard confidence level in media is 95%?


HOW CAN WE FIND OUT HOW MANY ARE HOLLOW? WE CAN'T OPEN THEM ALL!
I KNOW! TAKE A RANDOM SAMPLE AND ESTIMATE!