VŠB - Technical University of Ostrava
Faculty of Electrical Engineering and Computer Science
Department of Applied Mathematics

PROBABILITY AND STATISTCS
(Lecture 8: Random Samples, Point and Interval Estimations)

prof. Ing. Radim Briš, CSc.

http://www.am.vsb.cz/bris/

Teaching

Statistics 1.zip

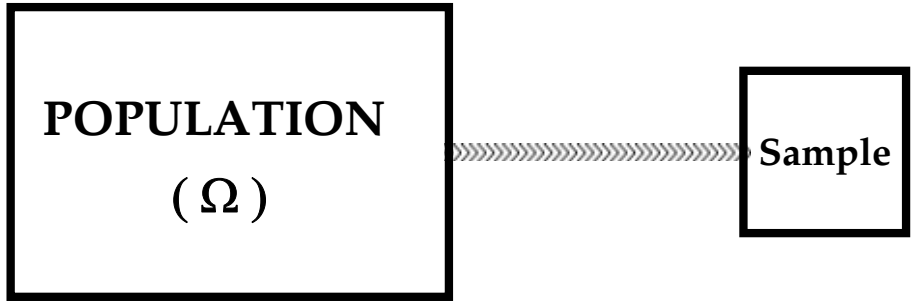
Random samples

Aim -

- understand the random sampling term
- use the sampling distribution and their properties

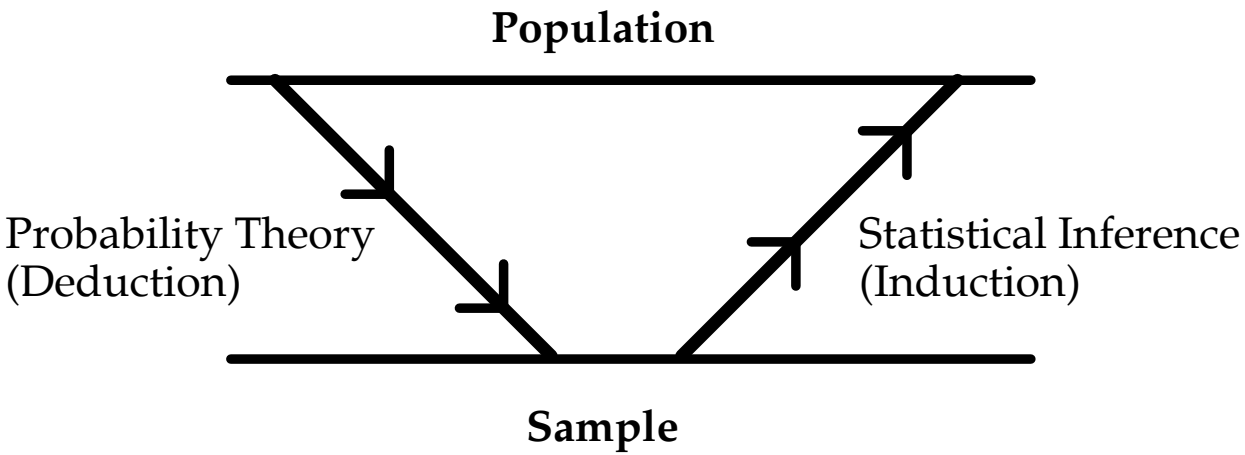
Population and Sample

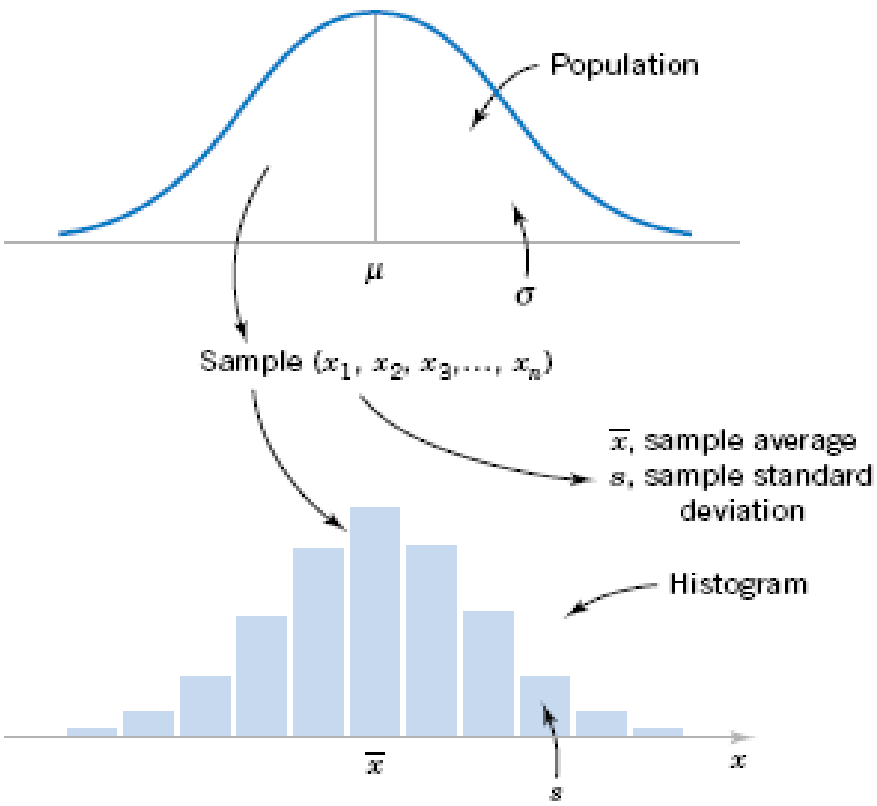
Deductive Theory of Probability



Known Probability Distribution
of Population

Sampling Distributions
of Sample Statistics





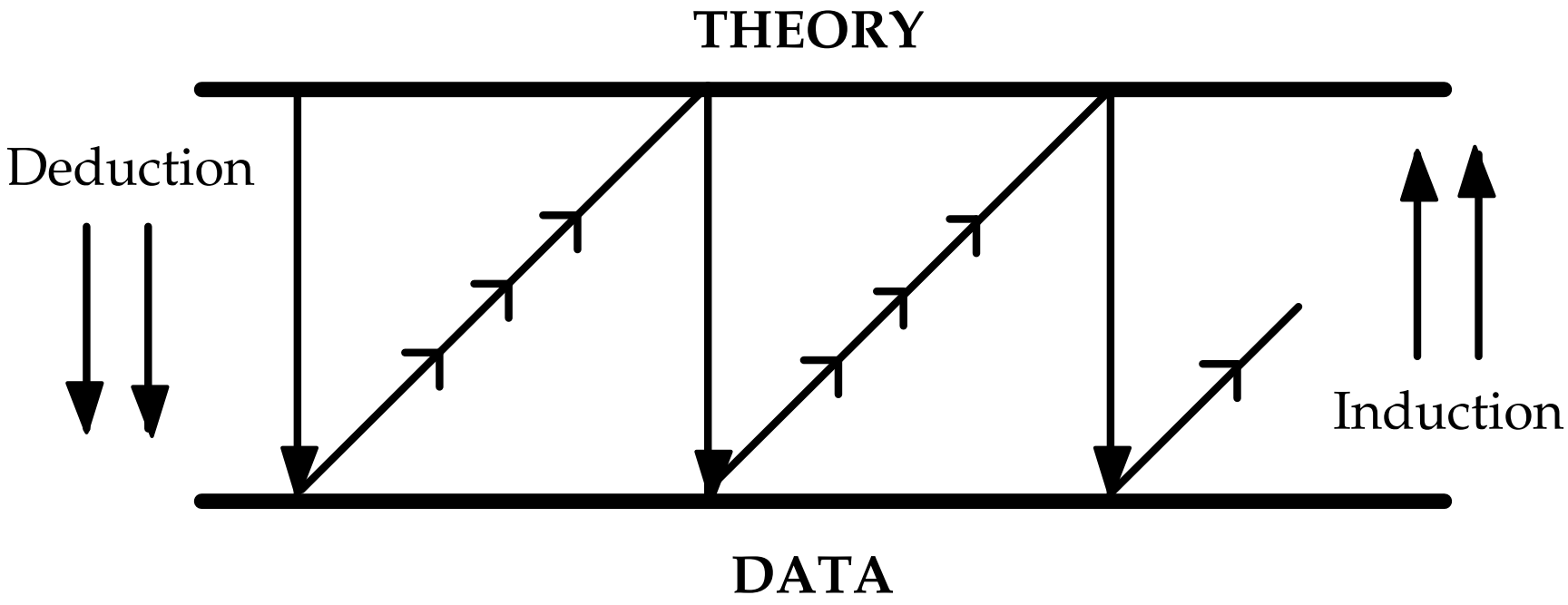
Definition:

A **population** consists of the totality of the observations with which we are concerned.

Definition:

A **sample** is a subset of observations selected from a population.

Deduction and induction



Statistical inference (induction) is the inverse of probability theory. It is the process of making statements about an unknown population on the basis of a known sample from that population.

Random Sample and Sampling Distributions

Random Sample $\underline{X}$ is special random vector: $\underline{X} = (X_1, \ldots, X_n)'$

Definition:

The random variables $X_1, X_2 \ldots X_n$ create a **Random Sample** $\underline{X} = (X_1, \ldots, X_n)'$ of size n if and only if:

(a) the X_i's are independent random variables, and

(b) every X_i has the same probability distribution.

Explanation:

Let X be a random variable that represents the result of one selection of an observation from the population. Let $f(x)$ denotes the probability density function of X.

Let us repeat observations, say n-times, each is obtained **independently**, under unchanging conditions.

We denote X_i as random variable that represents the i-th observation.

Then, $(X_1, \ldots, X_n)'$ is a random sample.

Obtained numerical values are usually denoted as $x_1, x_2, \ldots, x_n$.

Joint probability distribution of random sample

The random variables in a random sample are independent with the same probability distribution $f(x)$: i.e. the marginal probability density functions of $X_1, X_2 \ldots X_n$ are

$$f(x_1),\ f(x_2), \ldots f(x_n)$$

and **joint *pdf*** of random vector $\underline{X} = (X_1, \ldots, X_n)'$ must be $f(\underline{x}) = \prod_{i=1}^{n} f(x_i)$

By analogy, random vector has a **joint distribution function** $F(\underline{x})$:

$$F(\underline{x}) = F(x_1, \ldots, x_n) = P(X_1 < x_1, X_2 < x_2, \ldots) = F(x_1) \,.\, F(x_2) \ldots = \prod_{i=1}^{n} F(x_i)$$

Random Sample: $\underline{X}=(X_1, \dots, X_n)^T \quad X_i \sim F(x)$

Definition:

A **statistic** $T(\underline{X})$ is any function of the observations in a random sample.

$T(\underline{X})=T(X_1, \dots, X_n)$ is a random variable. Its observed value is $t=T(x_1,\dots, x_n)$

Most important statistics we already know from Lecture 1 - EDA:

1. $T_1(\underline{X}) = \overline{X} = \sum_{i=1}^{n} X_i / n$ ... **sample mean** (or average)

   $$E(T_1(\underline{X})) = E\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} EX_i = \frac{EX_i}{n} \cdot \sum_{i=1}^{n} 1 = EX_i$$

2. $T_2(\underline{X}) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2$ ..... **sample variance** ($=S^2$)

   We can easily proof that $ES^2 = DX_i$

3. $T_3(\underline{X}) = \sqrt{S^2} = S$ ... **sample standard deviation**

Sampling distributions

Let's assume that given random sample comes from normal distribution:

$\underline{X}=(X_1, \dots , X_n)'$, $X_i \to N(\mu,\sigma^2)$

1. $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n} \to N\left(\mu, \frac{\sigma^2}{n}\right)$ … comes from CLTfor large number n

2. $Z_n = \frac{\overline{X_n} - \mu}{\sigma} \cdot \sqrt{n} \to N(0,1)$ … comes from a transformation of previous distrib.

3. $\frac{S_n^2}{\sigma^2} \cdot (n-1) \to \chi^2(n-1)$ … was explained in context with χ^2discussion:

$$S_n^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2 \quad ; \frac{S_n^2}{\sigma^2} \cdot (n-1) = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2$$

Sampling distributions

4. $\dfrac{\overline{X}_n - \mu}{S} \cdot \sqrt{n} \to t_{n-1}$

… was derived in the discussion about using of Student's distribution, since:

$$\frac{\dfrac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n}}{\sqrt{\dfrac{\dfrac{S^2}{\sigma^2} \cdot (n-1)}{n-1}}} = \frac{\overline{X}_n - \mu}{S} \cdot \sqrt{n}$$

Assume two samples from the normal distribution

$\underline{X}=(X_1, \dots, X_n)'$, $X_i \to N(\mu_1, \sigma_1^2)$, $\underline{Y}=(Y_1, \dots, Y_m)'$, $Y_j \to N(\mu_2, \sigma_2^2)$. Then it holds:

5. $$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \to N(0,1).$$

Because:

$$\overline{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right)$$

$$\overline{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

$$\overline{X}-\overline{Y} \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}\right)$$

Two samples from the normal distribution - continuation

6. $$\frac{\dfrac{\dfrac{S_x^2}{\sigma_1^2}\cdot(n-1)}{n-1}}{\dfrac{\dfrac{S_y^2}{\sigma_2^2}\cdot(m-1)}{m-1}} = \frac{\dfrac{S_x^2}{\sigma_1^2}}{\dfrac{S_y^2}{\sigma_2^2}} \to F_{n-1,m-1}$$ … explained in F – distribution

Now assume that the variances are the same and unknown: $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Then can be shown that it holds:

7. $$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_x^2(n-1) + S_y^2(m-1)}} \cdot \sqrt{\frac{n.m.(n+m-2)}{n+m}} \to t_{n+m-2}$$

Two samples from the normal distribution - continuation

Explanation:

Let $Z_1 = \frac{S_x^2}{\sigma^2} \cdot (n-1) \to \chi^2(n-1)$ a $Z_2 = \frac{S_y^2}{\sigma^2} \cdot (m-1) \to \chi^2(m-1)$. Then

$$Z = Z_1 + Z_2 = \frac{S_x^2}{\sigma^2} \cdot (n-1) + \frac{S_y^2}{\sigma^2} \cdot (m-1) \to \chi^2(n+m-2)$$

Let:

$U = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \to N(0,1)$. Now we establish to t :

$$t = \frac{U}{\sqrt{\frac{Z}{m+n-2}}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_x^2(n-1) + S_y^2(m-1)}} \cdot \sqrt{\frac{n.m.(n+m-2)}{n+m}} \to t_{m+n-2}, \quad \ldots c.b.d.$$

These sampling distributions are very useful to construct important hypothesis tests as well as interval estimations, etc.

Point and interval estimation

Aim -

- explain the properties of the point estimation
- construct interval estimations for mean, standard deviation and variance
- examples

Point and interval estimation

(Problem of statistical inference)

From a methodological point of view we use two kinds of parameter estimations. It is a **point estimation** where distribution parameter is approximated by a number and so called **interval estimation** where this parameter is approximated by an interval where the parameter belongs with a high probability.

Basic notions

Let $(X_1, \dots, X_n)'$ be a random sample from probability distribution $F(x,\theta)$, θ is unknown parameter.

Parameter space … set of all possible values of θ .

Statistics $\hat{\theta} = T(\underline{X}) = T(X_1, \dots, X_n)$, which serves to estimate θ, is called **estimator** of θ.

An **estimate** is the value of that statistic for a particular sample result.

Properties of Estimators: unbiased estimator

Estimator $\hat{\theta}$ is **unbiased**, if $E\hat{\theta} = \theta$

If at least: $\lim_{n \to \infty} E\hat{\theta} = \theta$,

$\rightarrow$ **asymptotically unbiased estimator**

Examples:

Let $(X_1, \ldots, X_n)'$ be a random sample from normal distribution, let $\hat{\mu} = T(\underline{X}) = \overline{X}$.

Because $E\overline{X} = E(X_i) = \mu$,

$\rightarrow$ the sample mean is unbiased estimator of the parameter μ (which is expected value of the normal distribution at the same time).

Let $\hat{\sigma}^2 = T(\underline{X}) = S^2$ is sample variance.

Because we could proof: $E(S^2) = \sigma^2$,

$\rightarrow$ statistic S^2 is unbiased estimator of parameter σ^2 (which means variance for the normal distribution).

Properties of Estimators: efficiency

An estimator should be close to the parameter being estimated. Simply being unbiased will not insure that the estimator is close to the parameter. The variance of the sampling distribution of the estimator must be small as well.

If two estimators are unbiased, then the estimator with smaller variance is more efficient.
If an unbiased estimator has minimum possible variance for all unbiased estimators, then it is said to be **efficient**.

Properties of Estimators: consistency

Consistency is generally agreed to be an essential characteristic of an estimator.

Estimator $\hat{\theta}$ is **consistent** if and only if

a) $\hat{\theta}$ is asymptotically unbiased, tj. $E\hat{\theta} \to \theta$

b) $\lim\limits_{n \to \infty} D\hat{\theta} = 0$

Example:

$\overline{X}$ is consistent estimator of expected value, because

$$D\overline{X} = \frac{\sigma^2}{n} \to 0 \quad \text{for} \quad n \to \infty .$$

If $(X_1, \ldots, X_n)'$ is random sample from normal distribution, then:

(i) $\overline{X}$ is efficient and unbiased estimator of μ

(ii) S^2 is efficient and unbiased estimator of σ^2

Interval estimation: two sided estimation

- We have to find two statistics $T_D(\underline{X}), T_H(\underline{X})$ so that for any $\alpha > 0$, close 0, to be valid:

$$P(T_D(\underline{X}) \geq \theta) \leq \frac{\alpha}{2} \quad \text{a} \quad P(T_H(\underline{X}) \leq \theta) \leq \frac{\alpha}{2}$$

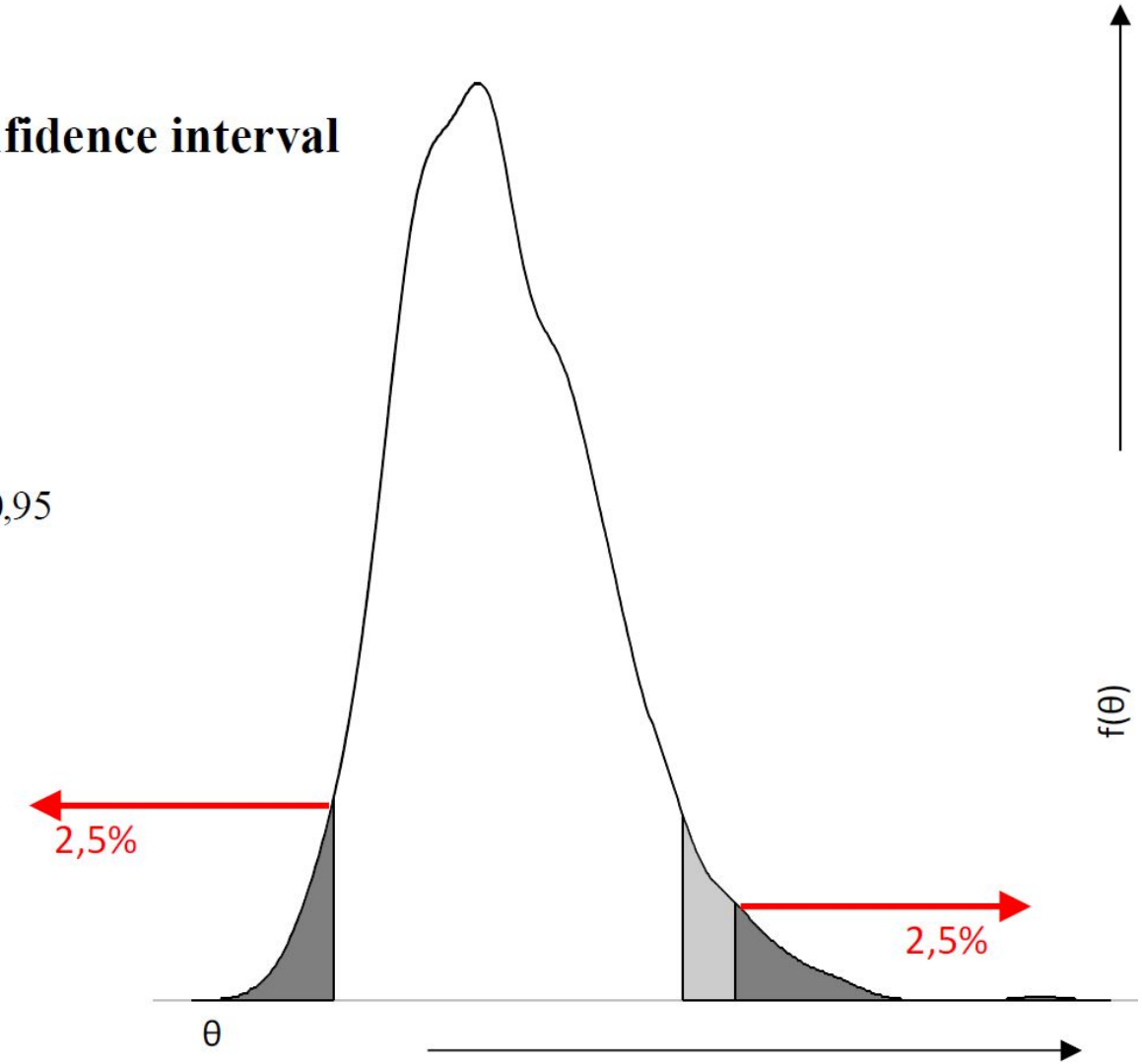
We call the value 1-α the **confidence level** of the interval.

These conditions result in $P(T_D(\underline{X}) \leq \theta \leq T_H(\underline{X})) \geq 1 - \alpha$.

Definition: The interval (T_D, T_H) we call **100.(1-α) % confidence interval** of parameter θ.

Two sided confidence interval

$$P(T_D \leq \theta < T_H) = 0{,}95$$



Interval estimation – one sided estimations

- We have to find one statistics $T_D(\underline{X})$, so that for any $\alpha > 0$, close 0, to be valid:

$$P(T_D(\underline{X}) \geq \theta) \leq \alpha$$

i.e.

$$P(T_D(\underline{X}) \leq \theta \leq +\infty) \geq 1 - \alpha$$

We speak about **one-sided confidence intervals**:

$$(T_D, +\infty)$$

Construction of Confidence Intervals – two Steps

Step 1: Select proper sampling distribution: $T(\underline{X}) \approx F(x)$

and find x_D x_H as the corresponding quantiles: $P(x_D < T(\underline{X}) < x_H) = 1-\alpha$

$$P(T(\underline{X}) < x_D) = P(T(\underline{X}) \geq x_H) = \frac{\alpha}{2}$$

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

$$x_D = F^{-1}\left(\frac{\alpha}{2}\right) = x_{\frac{\alpha}{2}}$$

$$x_H = F^{-1}\left(1-\frac{\alpha}{2}\right) = x_{1-\frac{\alpha}{2}}$$

$$P\left(x_{\frac{\alpha}{2}} < T(\underline{X}) < x_{1-\frac{\alpha}{2}}\right) = 1-\alpha \quad (*)$$

Step 2: Make final algebraic adjustments of (*) to find (T_D, T_H):

$$P(T_D(\underline{X}) \leq \theta \leq T_H(\underline{X})) \geq 1-\alpha$$

Confidence intervals for parameters of normal distribution

$$\underline{X} = \left(X_1, \ldots, X_n\right)^T$$

$$X_i \to N(\mu, \sigma^2)$$

Example 1

Let $(X_1, \dots, X_n)'$ is a random sample from normal distribution $N(\mu, \sigma^2)$.
Let σ^2 is known. Find confidence interval for μ: (T_D, T_H)
For a given small α.

Solution :

Step 1: We know (see Sampling distributions), that:

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n} \to N(0,1)$$

Step 2: we will find quantiles of the normal distribution: $u_{\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}$

$$P(u_{\frac{\alpha}{2}} \le \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \le u_{1-\frac{\alpha}{2}}) = 1-\alpha, \ (*)$$

Using symmetry of normal distribution: $u_{\frac{\alpha}{2}} = -u_{1-\frac{\alpha}{2}}$ and making algebraic adjustments of (*) we finally obtain:

Example 1 - continuation

$$P(\overline{X_n} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X_n} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha,$$

So that the interval (T_D, T_H) is found:

$$(T_D, T_H) = (\overline{X_n} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \quad \overline{X_n} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}})$$

Note:

Notice, that with increasing sample size *n* the width of the interval decreases so that confidence interval is more and more precise (at constant α). On the contrary having constant sample size *n*, with increasing confidence level 1-α the interval width increases.

Example 2

Suppose the same example but variance σ^2 is not known now.

Solution:

Step 1: We know (see Sampling distributions), that:

$$\frac{\overline{X}_n - \mu}{S} \cdot \sqrt{n} \to t_{n-1}$$

Step 2: We will find quantiles of the Student distribution (with *n-1* degrees of freedom): $t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}$

$$P(t_{\frac{\alpha}{2}} \le \frac{\overline{X}_n - \mu}{S}\sqrt{n} \le t_{1-\frac{\alpha}{2}}) = 1-\alpha, \ (*)$$

Using symmetry of Student distribution: $t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$ and making algebraic adjustments of (*) we finally obtain (T_D, T_H) : $P(\overline{X}_n - t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \le \mu \le \overline{X}_n + t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}) = 1-\alpha$

Example 3

Let $(X_1, \dots, X_n)'$ is a random sample from normal distribution $N(\mu, \sigma^2)$.
Find confidence interval for variance σ^2 : (T_D, T_H)
For a given small α .

Solution:

Step 1: We know (see Sampling distributions), that: $\frac{S_n^2}{\sigma^2} \cdot (n-1) \to \chi^2(n-1)$

Step 2: we will find quantiles of the Chi-square distribution: $\chi^2_{\frac{\alpha}{2}}(n-1), \chi^2_{1-\frac{\alpha}{2}}(n-1)$

$$P(\chi^2_{\frac{\alpha}{2}}(n-1) \le \frac{S_n^2}{\sigma^2} \cdot (n-1) \le \chi^2_{1-\frac{\alpha}{2}}(n-1)) = 1-\alpha \qquad (*)$$

making algebraic adjustments of (*), we finally obtain:

$$P\left(\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} S_n^2 \le \sigma^2 \le \frac{n-1}{\chi^2_{\frac{\alpha}{2}}(n-1)} S_n^2\right) = 1-\alpha$$

so that the confidence interval is: $\left(\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} S_n^2, \ \frac{n-1}{\chi^2_{\frac{\alpha}{2}}(n-1)} S_n^2\right)$

Example 3 - continuation

Having the confidence interval for variance

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

We can easily find (by square root) the confidence interval for standard deviation, for the same α:

$$\left(S_n \cdot \sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}},\ S_n \cdot \sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2}}(n-1)}}\right)$$

Example 4

Let us suppose a market investigation: 180 people questioning, results show mean consumption of fast food per one person per week: $(0{,}82 \pm 0{,}48)$.
Find 95% confidence interval of mean consumption of all population (in given town).

Solution:

$$\left.\begin{array}{l} \overline{x}_n = 0{,}82 \\ s = 0{,}48 \\ n = 180 \end{array}\right\} \Rightarrow \text{we select Student sample statistics (}\textbf{\textit{Step 1}}\text{):} \ \frac{\overline{X}_n - \mu}{S} \cdot \sqrt{n} \to t_{179}$$

Step 2: $$P(t_{0,025} < \frac{\overline{X}_n - \mu}{S}\sqrt{n} < t_{0,975}) = 0{,}95.$$

Example 4 - continuation

Seeing a Tables of Student distribution or STATGRAPHICS as well, we know that:
$t_{0,975} = -t_{0,025} = 1.96$,

where $t_{\frac{\alpha}{2}} = t_{0,025}$,$t_{1-\frac{\alpha}{2}} = t_{0,975}$ are quantiles of Student distribution with 179 degrees of freedom.

$$P(-1{,}96.\frac{S}{\sqrt{n}} - \overline{X}_{180} < -\mu < 1{,}96.\frac{S}{\sqrt{n}} - \overline{X}_{180}) = 0{,}95$$

$$P(\overline{X}_{180} - 1{,}96.\frac{S}{\sqrt{n}} < \mu < \overline{X}_{180} + 1{,}96.\frac{S}{\sqrt{n}}) = 0{,}95$$

Finally we can establish given data to obtain 95% confidence interval of mean consumption μ:

$$\boldsymbol{\mu \in (0{,}75;\ 0{,}89)}$$

Interval estimation of sample proportion: one proportion π

The same procedure as above, i.e.

Step 1: we know from CLT that $$P_1 = \frac{p-\pi}{\sqrt{\pi(1-\pi)}} \cdot \sqrt{n}; \quad \mathrm{P}_1 \to N(0;1)$$

Step 2:
Confidence interval is easy to find:

$$P\left(p - \sqrt{\frac{p\cdot(1-p)}{n}} \cdot z_{1-\frac{\alpha}{2}} < \pi < p + \sqrt{\frac{p\cdot(1-p)}{n}} \cdot z_{1-\frac{\alpha}{2}} \right) = 1-\alpha$$

Interval estimation of sample proportion: difference of two proportions $\pi_1 - \pi_2$

Step 1: we know from CLT that

$$P_2 = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}; \qquad \mathrm{P}_2 \to N(0;1)$$

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

Step 2:
Confidence interval is easy to find:

$$P\left((p_1 - p_2) - \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \cdot z_{1-\frac{\alpha}{2}} < (\pi_1 - \pi_2) < (p_1 - p_2) + \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \cdot z_{1-\frac{\alpha}{2}}\right) =$$

$$= 1 - \alpha$$

Example 5

Flash disks of two producers: Sonik a 5M underwent quality test:
600 tested disks of Sonic: 24 failed disks
500 tested disks of 5M: 14 failed disks.
Find 95% confidence interval for difference of failed disks Sonic and 5M.

Solution:

$$P\left((p_1 - p_2) - \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \cdot z_{1-\frac{\alpha}{2}} < (\pi_1 - \pi_2) < (p_1 - p_2) + \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \cdot z_{1-\frac{\alpha}{2}} \right) =$$

$$= 1 - \alpha$$

Example 5 - continuation

$x_1 = 24$ $\qquad$ $x_2 = 14$

$n_1 = 40 \cdot 20 = 800$ $\qquad$ $n_1 = 30 \cdot 20 = 600$

$p_1 = \frac{24}{800} = 0{,}030 \quad proportion_Sonic$ $\qquad$ $p_1 = \frac{14}{600} = 0{,}023 \quad proportion_5M$

$$p = \frac{24+14}{800+600} = 0{,}027$$

$$\frac{\alpha}{2} = 0{,}025; \quad 1 - \frac{\alpha}{2} = 0{,}975 \qquad \Rightarrow \qquad z_{0{,}975} = 1{,}96$$

Finally we can find: $\qquad P\left(-1{,}0\ \% < \left(\pi_1 - \pi_2\right) < 2{,}4\ \%\right) = 0{,}95$