Politechnika Gdańska realizuje projekt dofinansowany z Funduszy Europejskich „Zintegrowany Program Rozwoju Politechniki Gdańskiej”

Celem projektu jest podniesienie jakości kształcenia na studiach II i III stopnia, zwiększenie efektywności zarządzania Politechniką Gdańską oraz podniesienie kompetencji kadr.

Dofinansowanie projektu z UE: 28 905 073,51 zł
POWR.03.05.00-00-Z044/17

Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Społeczny

Statistics

Wojciech Artichowicz, PhD. Eng.

wojartic@pg.edu.pl

Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Społeczny

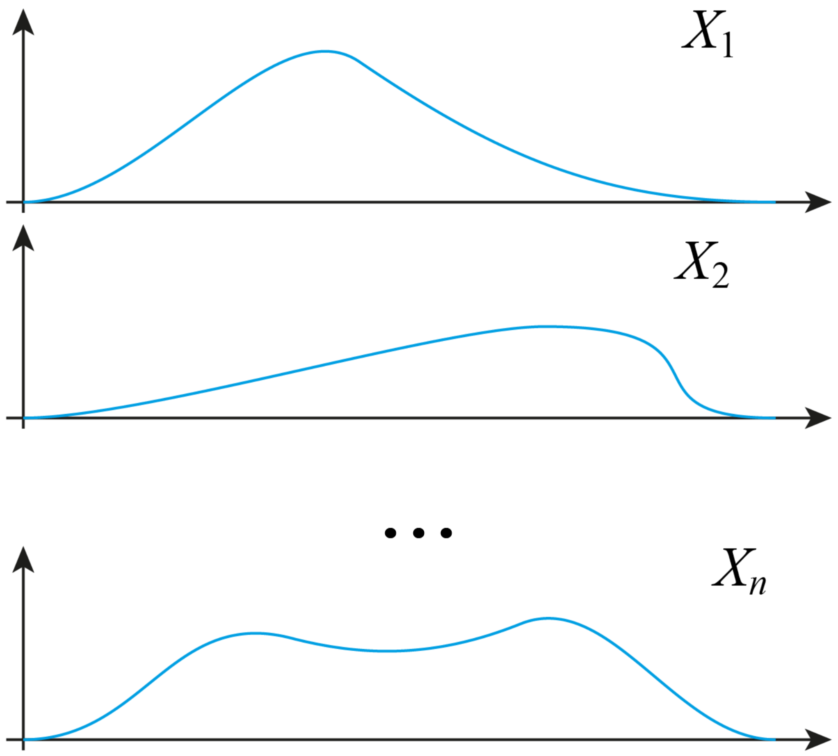
Important theorems

Series of random variables

$\{X_n\}$ is a series of independent random variables, if for any $n \geq 2$ random variables $X_1, X_2, \ldots, X_n$ are independent.

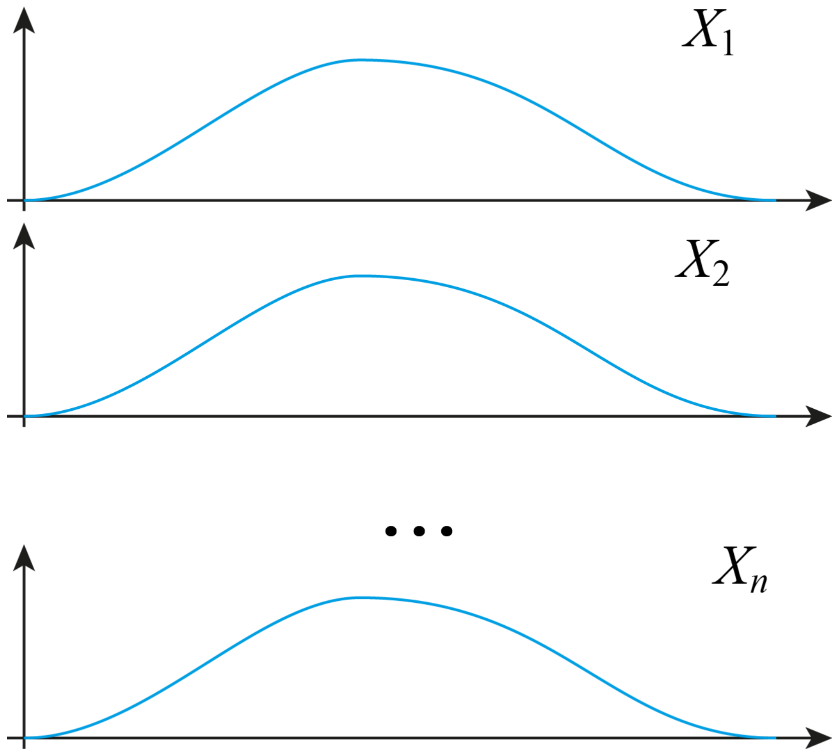
Series of random variables (example)

Series of random variables with different probability distributions



Series of random variables (example)

Series of random variables with identical probability distributions



(Weak) convergence of series of cumulative distribution functions

Series of CDFs $F_1(x)$, $F_2(x)$, ... is convergent to a non-decreasing function $F(x)$, if for every continuity point x holds

$$\lim_{n \to \infty} F_n(x) = F(x)$$

$F(x)$ is called the limiting cumulative distribution function, whereas $f(x)$ is limiting (or asymptotic) dstribution.

Convergence of series of random variables

Series of RVs X_1, X_2, ... is stochastically convergent to random variable X if for any positive small number holds

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$$

If $X = C = const$. then the series of random variables is stochastically convergent to a constant C.

For $Y_n = X_n - X$ the series of random variables $\{Y_n\}$ is stochastically convergent to zero if holds

$$\lim_{n\to\infty} P(|Y_n| > \varepsilon) = 0$$

Limit theorems

Limit theorems concern the limiting properties of series of random variables.

Limit theorems:

- local – consider probability mass functions (discrete RV) or probability density functions
- integral – consider cumulative probability distribution functions
- large number theorems – consider stochastic convergence to zero of some RVs

Central limit theorems

Limit theorems concerning the sums of independent random variables which limiting CDF is normal distribution CDF are called central limit theorems.

Some limit theorems

Integral	Local	Large number
Lindberg-Levy’s theorem	Poisson’s theorem	Chinchin law
Moivre-Laplace central limit theorem	Moivre-Laplace theorem	Bernoulli law
Lapunov theorem		Borel-Cantelli law

Bienaymé-Chebyshev theorem

If random variable X has finite expected value μ and variance σ^2, then for every positive number ε holds

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

Explanation:

The probability that RV X will take a value which differs from the expected value not less than ε is not greater than $\frac{\sigma^2}{\varepsilon^2}$

Bienaymé-Chebyshev theorem (rearrangement)

Substituting $\varepsilon = k \cdot \sigma$ ($k \in \mathbb{R}$) inequality becomes

$$P(|X - \mu| < k \cdot \sigma) \geq 1 - \frac{1}{k^2}$$

Explanation:

The probability that event $|X - \mu| < k \cdot \sigma$ will occur is greater than $1 - \frac{1}{k^2}$

remark:

$$|X - \mu| < k \cdot \sigma$$
$$-k \cdot \sigma < X - \mu < k \cdot \sigma$$
$$\mu - k \cdot \sigma < X < \mu + k \cdot \sigma$$

Bienaymé-Chebyshev theorem (rearrangement)

If considered probability distribution is continuous and unimodal inequality becomes

$$P(|X - \mu| < k \cdot \sigma) \geq \frac{4}{9} \cdot \frac{1 + s^2}{(k - |s|)^2}$$

for $k > |s|$, where $s = \frac{\mu - D}{\sigma}$, D – modal value.

Moivre-Laplace integral theorem

Let $\{X_n\}$ be a series of random variables with binomial distribution

$$P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

with $E[X_n] = n \cdot p$ and $V[X_n] = n \cdot p \cdot (1-p)$.

If $\{F_n(y)\}$ is a series of CDFs of standardized random variables

$$Y_n = \frac{X_n - n \cdot p}{\sqrt{n \cdot p \cdot q}}$$

then for each value of y holds

$$\lim_{n\to\infty} F_n(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{1}{2}u^2} du$$

Linberg-Levy central integral theorem

If $\{X_n\}$ is a series of independent random variables with identical distributions with finite expected value μ and variance $\sigma^2 \neq 0$. Then for the series of CDFs $\{F_n(z)\}$ of random variable

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n \cdot \mu}{\sigma \cdot \sqrt{n}}$$

holds

$$\lim_{n \to \infty} F_n(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}u^2} du$$

Linberg-Levy central integral theorem

Lindberg-Levy theorem implies that the random variable

$$Y_n = \sum_{i=1}^{n} X_i$$

has normal distribution with expected value $E[Y_n] = n \cdot \mu$ and standard deviation $S[Y_n] = \sigma\sqrt{n}$

Linberg-Levy central integral theorem

Lindberg-Levy theorem implies that the random variable

$$U_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which is mean value of n independent random variables, has normal distribution with expected value $E[Y_n] = \mu$ and standard deviation $S[Y_n] = \frac{\sigma}{\sqrt{n}}$

Khinchin's large number theorem

If $\{X_n\}$ is a series of independent random variables with identical probability distribution and finite expected value $E[X_n] = \mu$, then the series $\{Y_n\}$ of random variables

$$Y_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

is stochastically convergent to constant μ. That is

$$\lim_{n \to \infty} P(|Y_n - \mu| \geq \varepsilon) = 0$$

where ε is a small positive number.

Statistical estimation

Statistical estimation

Inference about statistical population pdf, based on a sample.

Types of statistical estimation:

- parametric
- non-parametric

Parametric estimation

Sample based inference about the value or values of parameters, of general population probability distribution.

Types of parametric estimation:

- point estimation
- interval estimation
- hypotheses testing

Non - parametric estimation

Sample based inference about the general population without specyfing the parameters characterizing the general population.

Types of non-parametric estimation:

- distribution based
- distribution free

Parametric estimation

Estimand

Distribution parameter Θ of interest which is to be estimated is called an estimand.

Estimator of a distribution parameter

Estimator of a parameter Θ of a probability distribution of random variable X is a function of a sample, which is a random variable with probability distribution depending on Θ and inference about Θ can be based on this function.

Let $X_1, X_2, \dots, X_n$ be a series of random variables with CDF's $F_i(x; \Theta)$ observed in a sample. Estimator based on those random variables will be denoted as $\hat{\Theta}_n = \hat{\Theta}_n(X_1, X_2, \dots, X_n; \Theta)$.

Estimator of a distribution parameter

Statistic $\widehat{\Theta}_n(X_1, X_2, \ldots, X_n)$ used for estimating unknown value of the parameter of probability distribution is called an *estimator*.

Estimate

Explicit value of estimator is called an estimate.

Estimand, estimator, estimate (example)

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ , σ – distribution parameters, estimands

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$\hat{\mu}, \hat{\sigma}$ – estimators

$$m = 3$$
$$s = 1$$

m, s – estimates

Estimator properties

- Biasedness
- Consistency
- Efficiency
- Mean square error of an estimator

Unbiased estimator

Estimator $\hat{\Theta}_n(X_1, X_2, \ldots, X_n;\ \Theta)$ of parameter Θ, which is unbiased if

$$E\left[\hat{\Theta}_n\right] = 0$$

Estimator biasedness

Estimator $\hat{\Theta}_n(X_1, X_2, \ldots, X_n;\, \Theta)$ of parameter Θ, which is not unbiased is called biased one. *Bias* of an estimator is

$$B_n = E\left[\hat{\Theta}_n\right] - \Theta$$

Estimator asymptotically unbiased

Estimator $\hat{\Theta}_n(X_1, X_2, \ldots, X_n; \Theta)$ of parameter Θ, which is called asymptotically unbiased, if

$$\lim_{n\to\infty} B_n = 0$$

Unbiased estimator (example)

Test if sample mean is unbiased estimator of expected value of a probability distribution

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$E(X_i) = \mu, \text{ for } i = 1,2,\dots,n$$

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} \cdot n \cdot \mu = \mu$$

Consistient estimator

Estimator $\hat{\Theta}_n(X_1, X_2, \dots, X_n;\ \Theta)$ of parameter Θ, which is called consistent

$$\lim_{n\to\infty} P\left(\left|\hat{\Theta}_n - \Theta\right| < \varepsilon\right) = 1$$

for arbitrarily small ε.

If the variance of an estimator $\hat{\Theta}_n(X_1, X_2, \dots, X_n; \Theta)$ of parameter Θ has a limit zero with n going to infinity and this estimator is unbiased or symptotically unbiased then, it is also consisitent estimator.

Effectiveness of an estimator

Let be given the set $\mathcal{T}$ of estimators $\widehat{\Theta}_n^{(1)}$, $\widehat{\Theta}_n^{(2)}$, ..., $\widehat{\Theta}_n^{(k)}$ of one parameter Θ.

Effectiveness of an estimator $\widehat{\Theta}_n^{(i)}$ is a ratio of a variance of an estimator with the smallest variance to the varince of an estimator $\widehat{\Theta}_n^{(i)}$

$$e = \frac{V(\widehat{\Theta}_n^{(min)})}{V(\widehat{\Theta}_n^{(i)})}$$

The most effective estimator

Estimator $\hat{\Theta}_n^{(i)}$ is called the most effective estimator if

$$e = \frac{V(\hat{\Theta}_n^{(min)})}{V(\hat{\Theta}_n^{(i)})} = 1$$

Minimal variance unbiased estimator

Unbiased estimator $\hat{\Theta}_n^{(i)}$ of a parameter Θ is called *minimal variance unbiased estimator* if among all unbiased estimators of Θ its variance has the smallest value.

Estimator is called the most effective estimator if

$$V\left(\hat{\Theta}_n^{(i)}\right) \leq V\left(\hat{\Theta}_n^{(j)}\right)$$

for $j = 1, 2, \ldots$ and $j \neq i$.

Mean square error of an estimator

Mean value of the squared distance $\left(\hat{\Theta}_n - \Theta\right)^2$ is called square error of the estimator $\hat{\Theta}_n$.

For any estimator $\hat{\Theta}_n$ its mean square error (MSE) is a sum of its variance and squared bias

$$E\left[\left(\hat{\Theta}_n - \Theta\right)^2\right] = V\left(\hat{\Theta}_n\right) + B_n^2$$

Choosing an estimator

Properties which an estimator should have

- Unbiasedness or asymptotic unbiasedness
- Least variance
- Greatest effectiveness
- Least mean square error

Determining estimators

Methods of estimator finding

- method of moments
- maximum likelihood method
- method of linear moments
- least squares method

Point estimation

Point estimation relies on finding such value, which can be regarded as a best estimation, of searched parameter of a distribution of a random variable.

End of lecture 9