ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК

ЮЖНЫЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

Empirical Characteristics Of The Sample.

Theoretical Background

Kurbatova Natalia Victorovna

nvkurbatova@sfedu.ru

Random variable (r.v.). Sample, based on single elementary outcome.

Table of probabilities of ξ∗ (distribution density)

Let's introduce new random variable ξ∗, which takes values $x_1, x_2, \ldots, x_n$ on single elementary outcome as n-surfaces dice with the same probability.

Definition 1. On a random variable (r.v.)
Given Ω - sample space and probability distribution P. A random variable ξ is a function defined on Ω, taking values in R

$$\xi : \Omega \to R$$

(We think, probabilistic space (Ω,A,P) is given!)

ξ^*	x_1	x_2	$\ldots$	x_n
$\tilde{p}$	$1/n$	$1/n$	$\ldots$	$1/n$

The distribution function:

$$F_n^*(y) = \sum_{x_i < y} \frac{1}{n} = \frac{\text{number } x_i \in (-\infty, y)}{n}$$

Definition 2. About the outcome

The elementary events (outcomes) of a random experiment are such events (atoms of the events) which exactly one of them occurs during the experiment.

-- Distribution function for built ξ∗

Descriptive statistics of ξ^*

$$E\xi^* = \sum_{i=1}^{n} \frac{1}{n} x_i = \frac{1}{n} \sum x_i = \bar{x};$$ **Expectation** $$E(\xi^*)^k = \bar{x}^k;$$

Variance :

$$D\xi^* = \sum_{i=1}^{n} \frac{1}{n} (x_i - E\xi^*)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = S^2; \qquad (1)$$

$$E\,g(\xi^*) = \frac{1}{n} \sum_{i=1}^{n} g(x_i) = \overline{g(x)};$$

If we allow an elementary outcome to change – obtain r.v. ξ

$F_n^*(y)$, $\bar{x}$, S^2, $\bar{x}^k$, $\overline{g(x)}$ as functions of r.v. $x_1, \ldots, x_n$ instead (1).

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

F (y) = P (x1<y) - Empirical Distribution Function (EDF)

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Empirical distribution function determined on the sample $\boldsymbol{x} = (x_1, \dots, x_n)$ volume n – is random function $F_n^*(y): R \times \Omega \to [0,1]$, $\forall$ y ∈ R, such that

$$F_n^*(y) = \frac{number\ \ X_i \in (-\infty, y)}{n} = \frac{1}{n}\sum_{i=1}^{n} I(x_i < y);$$

Definition 1. EDF

$$I(x_i < y) = \begin{cases} 1, & if\ x_i < y \\ 0, & else \end{cases}$$ – indicator function.

Definition 2. EDF *(after reordering)*:

$x_{(1)} \le x_{(2)} \le \dots \le x_{(n)};\ x_{(1)} = \min_i\{x_i\}, x_{(n)} = \max_{(n)}\{x_i\}$

$x_{(k)} - k\text{-}th\ element, or\ k\text{-}th\ order\ statistics.$

$F_n^*(y)$ has jumps in each point x_i *with* height $\frac{m}{n}$,

m – number of elements are equal to x_i

$$F_n^*(y) = \begin{cases} 0, & y \le x_{(1)} \\ \dfrac{m}{n}, & y \in (x_{(m)}, x_{(m+1)}) \\ 1, & y > x_{(n)}. \end{cases}$$

Remark.

For any y indicator of event xi <y, has Bernouli [bəˈnjuːlɪ] distribution with p – parameter: p = P(xi<y) = F(y), r.v. I (xi<y) ∈ B(F(y)) ,

Lecture 2 § 3.

Empirical probability function (EPF)

- Histogram *of relative frequencies*
- Table *exists for discrete distributions*
- Density function – *for absolutely continuous distributions*

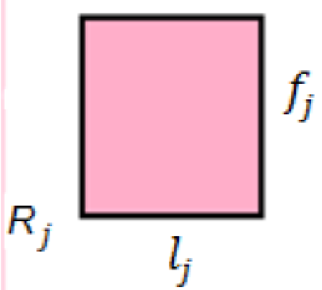
ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК

Histogram - empirical analogue of the density function of distribution

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Histogram is based on grouped data; the range of data values $x_{max} - x_{min}$ is devided into several bins $A_1, \dots, A_k$ (consecutive, non-overlapping intervals, often equal size - length): v_j – number elements of the sample got into j – th bin A_j. Let construct rectangle R_j with f_j – rectangle height and l_j – rectangle length (see pic.), where

$$f_j = \frac{v_j}{n * l_j}$$



$$v_j = \{\text{number } x_i \in A_j\} = \sum_{i=1} I(x_i \in A_j),$$

$$n = \sum_{j=1}^{k} v_j \quad for\ \forall A_j$$

$$f_j = \frac{v_j}{n * l_j} \text{ – analogue of the density function}$$

$$\frac{v_j}{n} \text{ – relative frequency}$$

Figure, consisting of union of k such rectangles, is called histogram, the total squares of rectangles is equal to 1 (S=1).

On relation between a histogram and a density function

Remark: Let density function of the sample, belongin to General population – continuous function. If number of intervals k=k(n) tends to infinity k (n)/ n→0, then histogram tends to density function for $\forall y$.

Task: Model such an example for any arbitrary distribution (Normal, Students, Chi-squared), shows the trend corresponding to the remark.

Sample and true moments

Remark. Let ξ – random variable belongs to General Population with true known characteristics; x_1 (or any x_i) – we can consider as outcome (result of experiment) of ξ and random variable too.

True moments	Estimation for true moments
$E\xi = \mathrm{E}x_1 = a$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;$
$D\xi = \mathrm{D}x_1 = \sigma^2$	$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ – biased ;
	$S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ – unbiased ;
$E\xi^k = \mathrm{E}x_1^k = m_k$	$S^k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ – k moments ;
$Eg(\xi)$	$\overline{g(x)} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Types of convergence . Definitions

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Definition 1: Sequence of r.v. $\{X_n\}$ almost probably converge to r.v. X, if

$$P\left\{\lim_{n\to\infty} X_n - X\right\} = 1.$$

Definition 2: Sequence of r.v. $\{X_n\}$ converges in probability to r.v. X, if

$$\lim_{n\to\infty} P\{|X_n - X| \le \varepsilon\} = 1.$$

Definition 3: Sequence of r.v. $\{X_n\}$ converges in the distribution to r.v. X (or weakly converges), if for all points of continuity of the probability distribution function the equality holds

$$\lim_{n\to\infty} F_{X_n}(x) = F_x(x) \quad \text{or} \quad F_{X_n}(x) \Longrightarrow F_x(x)$$

Remark 1: Convergence almost probably implies convergence in probability!

include

Remark 2: Convergence in probability implies convergence in distribution!

Convergens EDF To Theoretical One

Theorem 1. On convergence in probability EDF to theoretical distribution function.

Let $x_1, \dots, x_n$ – sample from distribution family $\mathcal{F}$ with distribution function F and let F_n^* – empirical distribution function, built in accordance with the sample, then

$$F_n^*(y) \xrightarrow{P} F(y), \text{ when } n \to \infty \text{ for } \forall y \in R.$$

Open Lecture 2 § 4. Have you questions?

Conclusion. With an increase in the sample size, the empirical distribution function tends in probability to an unknown theoretical distribution for $\forall y \in R$.

Repetition & Proof of Theorem1

Theorem 1 . On convergence in probability EDF to theoretical distribution function.

Proof

Using definition of EDF :

$$F_n^*(y) = \frac{number\ \ X_i \in (-\infty, y)}{n} = \frac{1}{n}\sum_{i=1}^{n} I(x_i < y).$$

Random variables $\mathrm{I}(x_1 < y)$, $\mathrm{I}(x_2 < y)$, …

$\underbrace{\text{independent and identically}}_{(*)}$ distributed =>

$$EI(x_1 < y) = 1 * P(x_1 < y) + 0 * P(x_1 \geq y) = P(x_1 < y) =$$

$$= F(y) < \infty;$$ *the consistency property of r.v.*

Repetition r.v. $\xi_1, \dots, \xi_n$: $E\xi_1 < \infty$, $E\xi_1 = a$

$$\frac{S_n}{n} = \frac{\xi_1 + \dots + \xi_n}{n} \to a;$$

according to LLN

$$F_n^*(y) = \frac{\sum_{i=1}^{n} I(x_i < y)}{n} \xrightarrow{\mathrm{P}} \mathrm{EI}(x_1 < y) = \mathrm{F}(y) \quad \square$$

Expectation of sum is equal sum of expectation for r.v. as ($*$)

Remark:

With increasing sample volume took from General Population, empirical distribution function tends in probability to unknown the same distribution for $\forall y \in R$

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Theorems of Convergence
(without proof)

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Theorem 2. (Glivenko–Cantelli)

In the condition of the theorem 1 we have: $\sup_{y \in R} |F_n^*(y) - F(y)| \underset{n \to \infty}{\to} 0$

Theorem 3. Kolmogorov

Let $x_1, \dots, x_n$ sample of F with continuous distribution function F, $F_n^*(y)$ – empirical distribution function for $\forall y \in R$, then

$$\sqrt{n} \sup_{y \in R} |F_n^*(y) - F(y)| \to \eta, \qquad n \to \infty$$

and η have Kolmogorov distribution with continuous function

$$F_\xi(y) = K(y) = \begin{cases} \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 y^2}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

On EDF Convergence

Theorem 4. (*on empirical distribution function*)

For any y ∈ R :

1) $EF_n^*(y)$= F(y), F_n^* – *unbiased estimation*
2) $DF_n^*(y) = \frac{F(y)(1-F(y))}{n}$;
3) $\sqrt{n}(F_n^*(y) - F(y)) \to N(0, F(y)(1 - F(y)))$ tends asymptotically F(y)≠0,1
4) $n\,F_n^*(y)$ *has* Binomial distribution $B(n, F(y))$

Proof:

Known: $I(x_i < y)$ have Bernoulli distribution

$I(x_i < y) \in B_{F(y)}$, =>

$EI(x_1 < y) = F(y)$, $DI(x_1 < y) = F(y)(1 - F(y))$

Theorem 4 Proof

1) $I(x_i < y)$ identically distributed, so $EF_n^*(y) = E\frac{\sum_{i=1}^n I(x_i<y)}{n} =$
$= \frac{nEI(x_1<y)}{n} = F(y)$

explain each step!

2) $I(x_i < y)$ – independent and identically distributed

$DF_n^*(y) = D\frac{\sum_{i=1}^n I(x_i<y)}{n} = \frac{\sum_{i=1}^n DI(x_i<y)}{n^2} = \frac{nDI(x_1<y)}{n^2} = \frac{F(y)(1-F(y))}{\boldsymbol{n}}$;

3) Based on central limit theorem

$\sqrt{n}(F_n^*(y) - F(y)) = \sqrt{n}\left(\frac{\sum_{i=1}^n I(x_i<y)}{n} - F(y)\right) = \left(\frac{\sum_{i=1}^n I(x_i<y) - nF(y)}{\sqrt{n}}\right) =$
$= \frac{\sum_{i=1}^n I(x_i<y) - nEI(x_1<y)}{\sqrt{\boldsymbol{n}}} \Rightarrow N(0, DI(x_1 < y)) = N(0, F(y)(1 - F(y));$

4) $nF_n^*(y) = I(x_1 < y) + \cdots + I(x_n < y) \in B(n, F(y))$;
(according to property: the stability of the summation)

Histogram properties

Let F – absolutely continuous distribution,

f – true density, k intervals of histogram

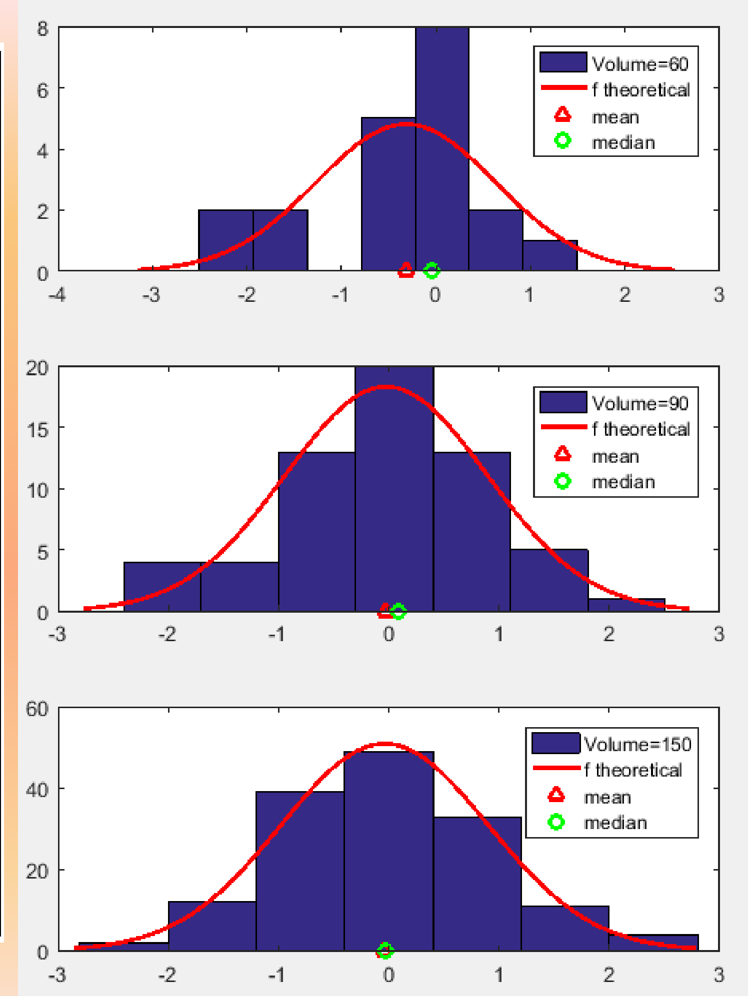
is not depends on n – sample volume.

With increasing n square of histogram column tends to square under density function.

Theorem 5: (without proof)

$$n \to \infty, \quad \text{for } \forall j = 1, \ldots, k;$$

$$l_j * f_j = \frac{\nu_j}{n} \to \mathrm{P}(x_1 \in A_j) = \int_{A_j} f(x)dx$$



Skewness. Characteristic of EDF symmetry

Skewness for univariate data $Y_1, Y_2, \ldots, Y_N,$ is:

$$g_1 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3/N}{s^3}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points.

Alternative definitions:

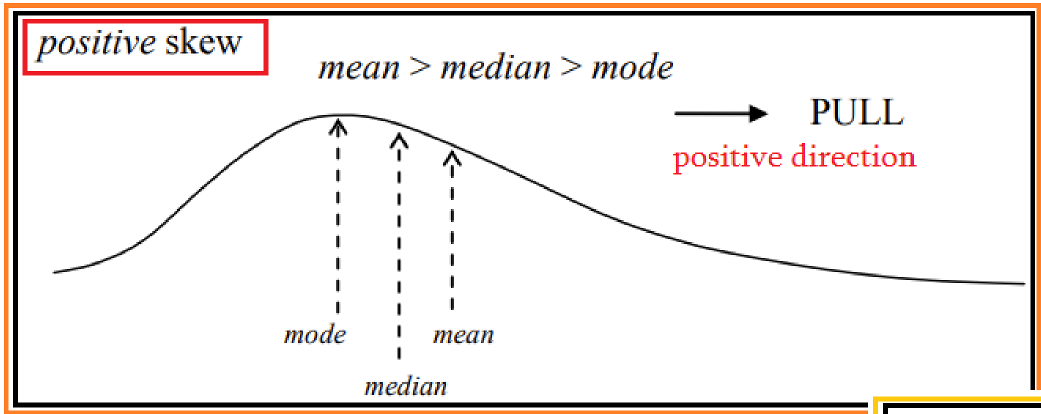
$$\text{Galton skewness} = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

here Q_1, Q_3 – lower and upper quartiles, Q_2 is the median.

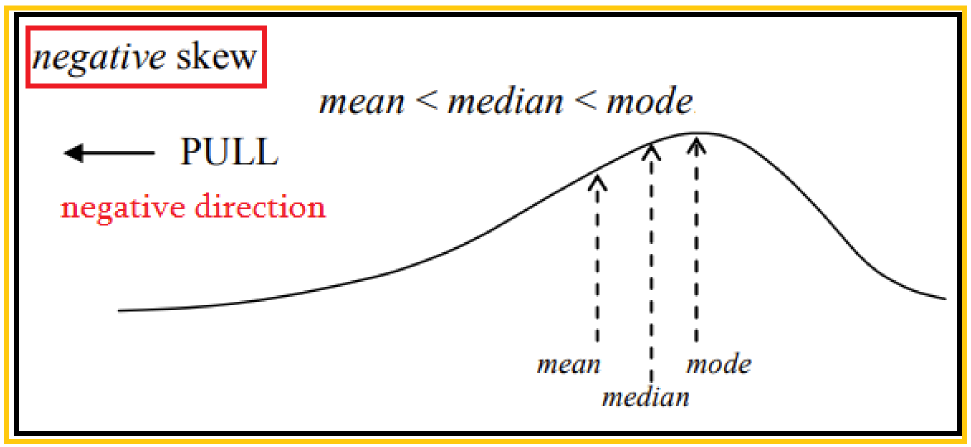
$$\text{Pearson2 skewness} = 3\frac{(\bar{Y} - \tilde{Y})}{s}$$

where $\tilde{Y}$ is the sample median.

Practical approach



See: Teams
“Practical manual.pdf”
pp. 16-21



<u>Correspondent MatLab functions:</u>

mean, median
modal - the most frequently value (existing depends on ML version)
var – variance , **std** – standard deviation
moment
skewness, kurtosis
histogram
histfit – histogram and density function of standard Normal distribution, density function is built with effective parameters

Moments of univariate data. Kurtosis (Excess)

Definition of Kurtosis **Kurtosis indicates how much data resides in the tails**

For univariate data $Y_1, Y_2, \ldots, Y_N$, the formula for kurtosis is:

$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points.

Alternative Definition of Kurtosis

The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis ("excess kurtosis"):

$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4} - 3$$

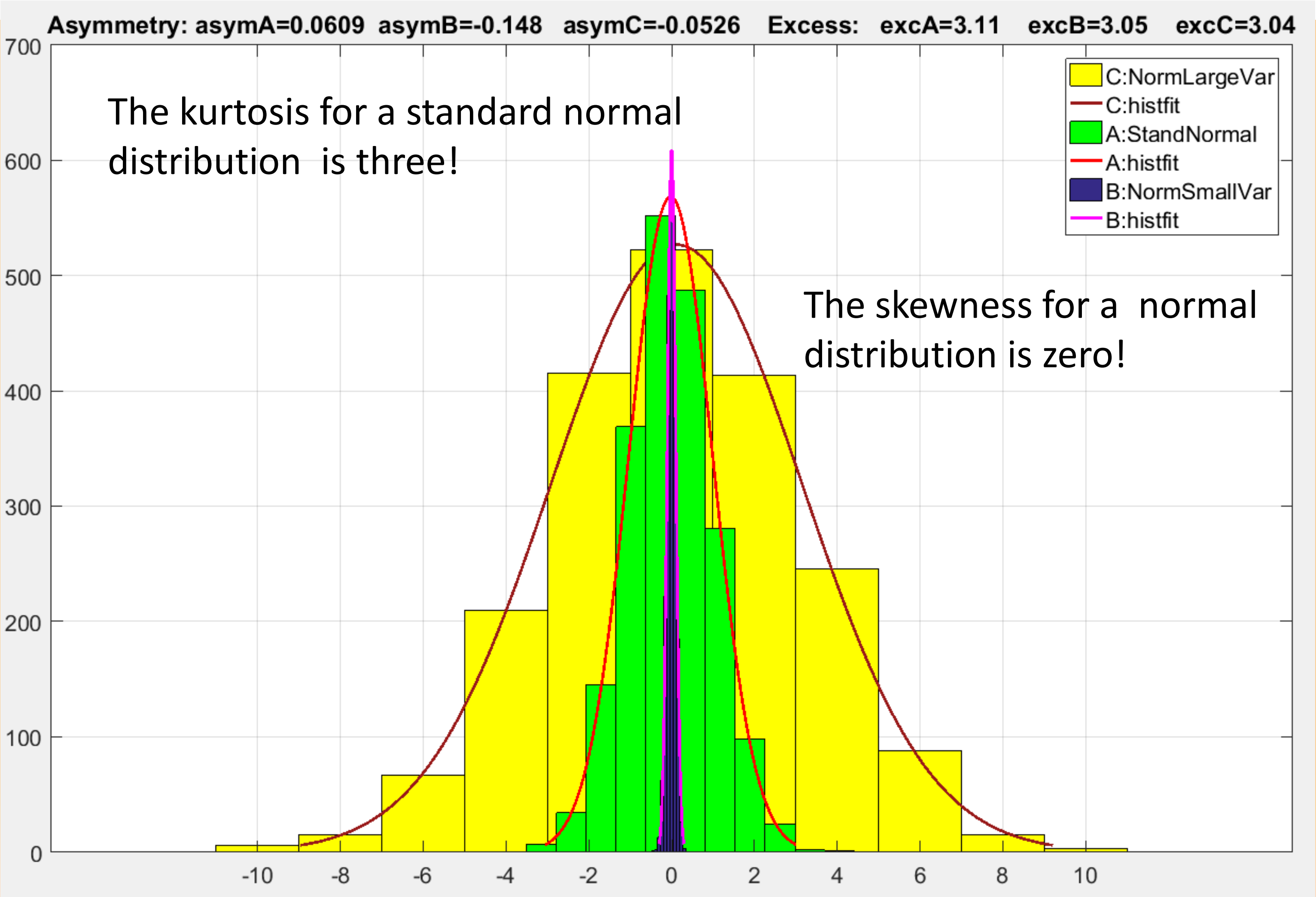
A large kurtosis – heavy data tails !

This definition is used so that the standard normal distribution has a kurtosis of three. In addition, with the second definition positive kurtosis indicates a "heavy-tailed" distribution and negative kurtosis indicates a "light tailed" distribution.

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Moments in Matlab

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

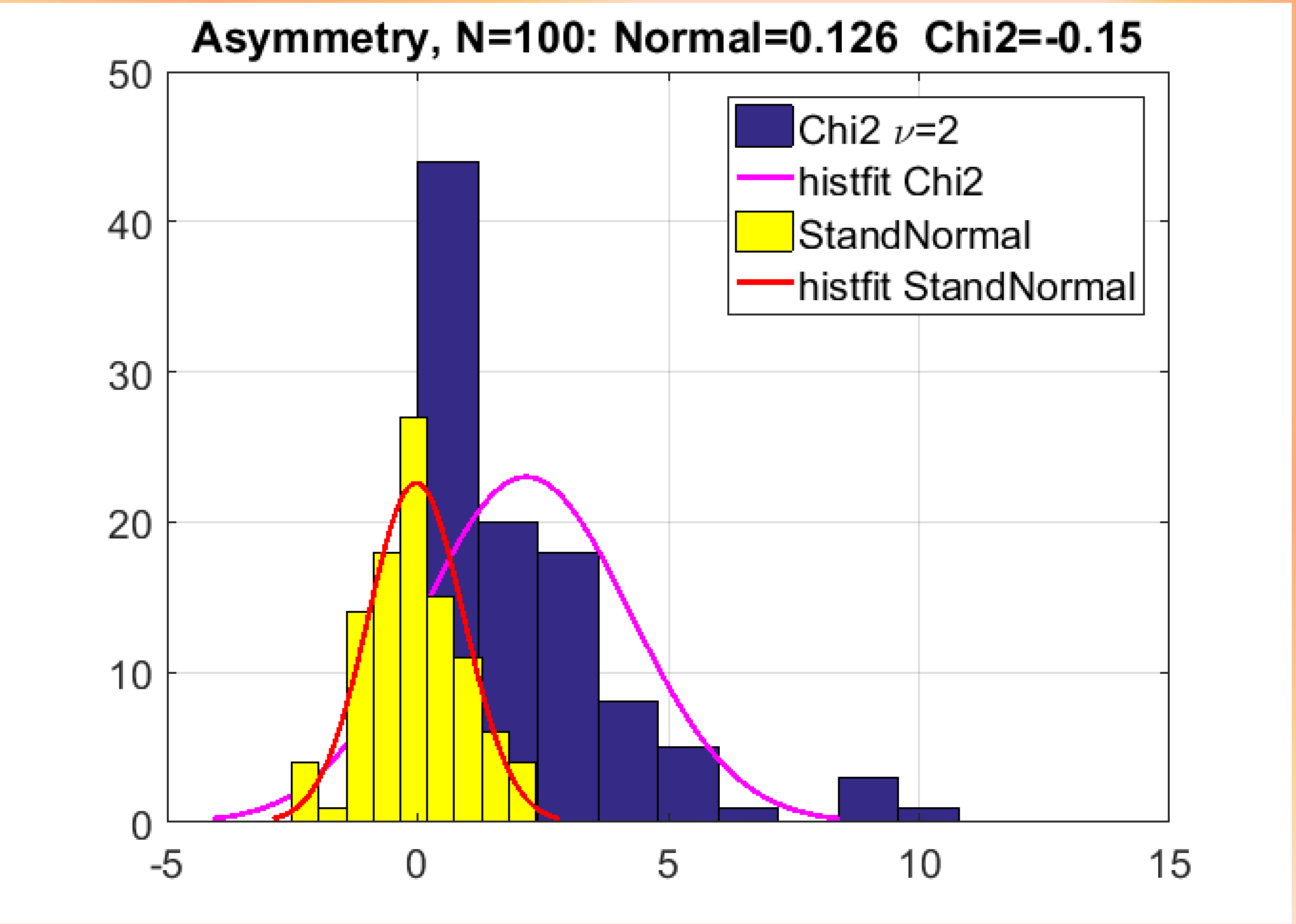


SkewnessKurtosis.m - script, corresponding to this plot. If you need, you can use Debug step by step!

ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК

Comparisen. Normal, Chi-Square

ЮЖНЫЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ



Theoretical Basis

What makes it possible to trust empirical characteristics?

Properties of the first-order moment

The sample mean $\bar{X}$ is unbiased, consistent and asymptotically normal estimation for the theoretical mean (expectation of r.v.)

Theorem 6

1) If $E|X_1| < \infty$, then $E\bar{X} = EX_1 = a$
2) If $E|X_1| < \infty$, then $\bar{X} \xrightarrow{P} EX_1 = a,\ n \to \infty$
3) If $D|X_1| < \infty,\ DX_1 \neq 0$, then $\sqrt{n}(\bar{X} - EX_1) \Rightarrow N(0,\ DX_1)$

($\bar{X}$ is an asymptotically normal estimate for the true expectation EX_1)

Proof: According to properties of expectation:

1) $E\bar{X} = \frac{1}{n}(EX_1 + .. + EX_n) = \frac{1}{n} nEX_1 = EX_1 = a - unbiased$
2) and LLN: $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{P} EX_1 = a$
3) $\sqrt{n}(\bar{X} - EX_1) = \frac{\sum_{i=1}^{n} X_i - nEX_1}{\sqrt{n}} = \frac{X_1 - EX_1 + \cdots + X_n - EX_n}{\sqrt{n}} = \ldots$

[according to (CLT): $\frac{X_1 - a + \cdots + X_n - a}{\sqrt{n}} \Rightarrow \mathcal{N}(0, \mathcal{D}X_1)$, $\mathcal{D}X_1 = \sigma^2$ – true variance],

that is if $X_i \in \mathcal{N}(a, \sigma^2)$ then $S_n = \frac{1}{n}\sum X_i \in \mathcal{N}\left(a, \frac{\sigma^2}{n}\right)$,

and as result we obtain $\ldots = \frac{X_1 - a + \cdots + X_n - a}{\sigma / \sqrt{n}} \in \mathcal{N}(0,1)$

ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК

Properties of the high-order moments

ЮЖНЫЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

Theorem 7

1. If $E|X_1|^k < \infty$, then $E\bar{X}^k = EX_1^k = m_k$
2. If $E|X_1|^k < \infty$, then $\bar{X}^k \xrightarrow{P} EX_1^k = m_k, \ n \to \infty$
3. If $DX_1^k < \infty$, and $DX_1^k \neq 0$ $then$ $\sqrt{n}\left(\bar{X}^k - EX_1^k\right) \Rightarrow \mathcal{N}\left(0, DX_1^k\right)$

Theorem 8 ***Variance properties.*** Let $DX_1 < \infty$, then

1. Sample variance $s^2 = \frac{1}{n}\sum_1^n (X_i - \bar{X})^2$ and $s_0^2 = \frac{1}{n-1}\sum_1^n (X_i - \bar{X})^2$ are consistent estimation for true variance: $s^2 \xrightarrow{P} DX_1 = \sigma^2$ and $s_0^2 \xrightarrow{P} DX_1 = \sigma^2$

2. Value s^2 – biased estimation of variance and s_0^2- unbiased one:

$$Es^2 = \frac{n-1}{n} DX_1 = \frac{n-1}{n}\sigma^2 \neq \sigma^2, \qquad Es_0^2 = DX_1 = \sigma^2$$

3. If $0 < D(X_1 - EX_1)^2 < \infty$, then s^2 and s_0^2 – asymptotically normal estimation of the true variance: $\sqrt{n}(s^2 - DX_1) \Rightarrow \mathcal{N}(0, D(X_1 - EX_1)^2)$

Typical Aproaches to proof

Proof:

1. $s^2 = \overline{X^2} - (\bar{X})^2 \xrightarrow{P} EX_1^2 - (EX_1)^2 = \sigma^2$. (theorem 7.1)

$$\frac{n}{n-1} \mapsto 1, \ so \quad s_0^2 = \frac{n}{n-1} s^2 \xrightarrow{P} \sigma^2$$

2. $ES^2 = E\left(\overline{X^2} - \bar{X}^2\right) =^{prop.E}= E\overline{X^2} - E(\bar{X}^2) = theor.7 = EX_1^2 - E(\bar{X}^2) =$ [т. к. $D(\bar{X}) = E\bar{X}^2 - (E\bar{X})^2$] $= EX_1^2 - ((E\bar{X})^2 + D(\bar{X})) = EX_1^2 - (EX_1)^2 - D\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \sigma^2 - \frac{1}{n^2} nDX_1 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2, \ ES_0^2 = \frac{n}{n-1} ES^2 = \sigma^2$

3. Introduce new variables $Y_i = X_i - a: \ DY_1 = DX_1 = \sigma^2, \quad EY_1 = 0$

Sample variance $s^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - a - (\bar{X} - a)\right)^2 = \overline{Y^2} - (\bar{Y})^2$,

so $\sqrt{n}(s^2 - \sigma^2) = \sqrt{n}\left(\overline{Y^2} - (\bar{Y})^2 - \sigma^2\right) = \sqrt{n}\left(\overline{Y^2} - EY_1^2\right) - \sqrt{n}(\bar{Y})^2 \ominus$

(theorem 7) $\ominus \left[\frac{\sum_{i=1}^n Y_i^2 - nEY_1^2}{\sqrt{n}} - \bar{Y}\sqrt{n}\bar{Y}\right] \Longrightarrow^{\{tends \ to\}} \mathcal{N}(0, DY_1^2) =$

$= \mathcal{N}(0, D(X_1 - a)^2)$, because: if $\bar{Y} \xrightarrow{P} EY_1 = 0$, then $\bar{Y}\sqrt{n}\bar{Y} \to 0$

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Conclusions

- $\overline{X}$ - first order momemt of the sample is unbiased (converge in averadge) and consistent (converge in probability)!
- $\overline{X^k}$ - k-order momemts (k>2) of the sample are unbiased and consistent!
- Both empirical variances $s_{0,}^2$ s^2 are consistent!
- But only s_0^2 - unbiased!
- Empirical moments are ASYMPTOTICALLY normal estimate of true moments.

THANKS FOR YOUR ATTENTION!
BE HEALTHY!