УНИВЕРСИТЕТ ИТМО

Methods & Models for Multivariate Data Analysis

Lecture 1. Univariate random variable analysis

Ass. Prof. Anna Kalyuzhnaya, PhD

Random Events Reminder

Backgrounds of random events theory

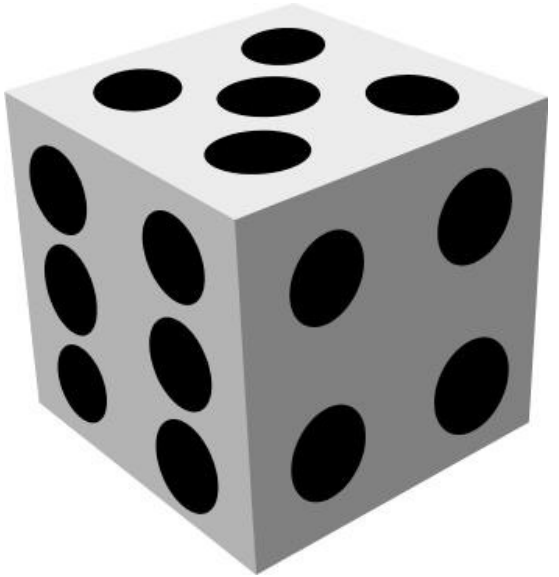
УНИВЕРСИТЕТ ИТМО

Виды событий: достоверное, невозможное, случайное.
Events are: reliable, impossible, random.

Reliable event – I
Impossible event - $\varnothing$

Random event – event that occurs with certain probability
Inverse (обратное) event – $\overline{A}$

Operations on events:
1. Sum - $\bigcup A_k$ **2. Production - $\bigcap A_k$**



Probability of events

УНИВЕРСИТЕТ ИТМО

Elementary event **is an outcome of experiment**

Random event **is a sum of set of elementary events** ω_i **:**

$$A = \cup\, \omega_i$$

All random events in a scheme form set of random events F – ***field or algebra of events*****.**

Axiomatic definition of probability:

Probability is a numerical function P(A) defined on F algebra, with properties:

1) **P(I)=1**
2) **P(Ø)=0**
3) **0 < P(A) < 1**

Probability of events

УНИВЕРСИТЕТ ИТМО

Classic definition of probability:

Probability of event is a ratio of number of outcomes (m) with event A to number of all outcomes (n) of experiments

$$P(A) = \frac{m}{n}$$

Statistical definition* of probability:

Probability of event is frequency of occurrence of this event that becomes stable in a case of large number of experiments

* Statistical definition doesn't give us formula for calculation of probability itself, but gives us opportunity for calculation an estimate of probability

$$P(A) \approx P^*(A) = \frac{\mu}{n}$$

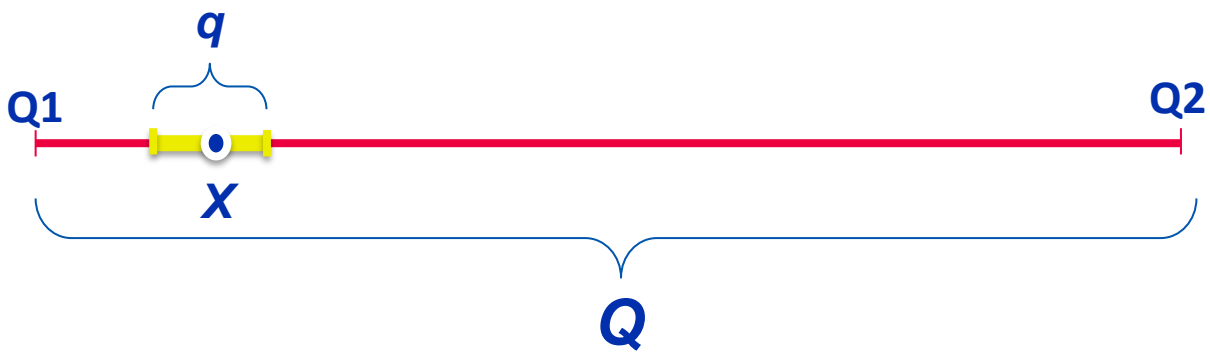
Probability of events

УНИВЕРСИТЕТ ИТМО

Geometric definition of probability:

Occurrence of X on the interval Q is a reliable event

Occurrence of X on the interval q is a random event



$$P(A) = \frac{length\ \ q}{length\ \ Q}$$

ITMO more than a UNIVERSITY

What are compatible and incompatible events?

Backgrounds of random events theory

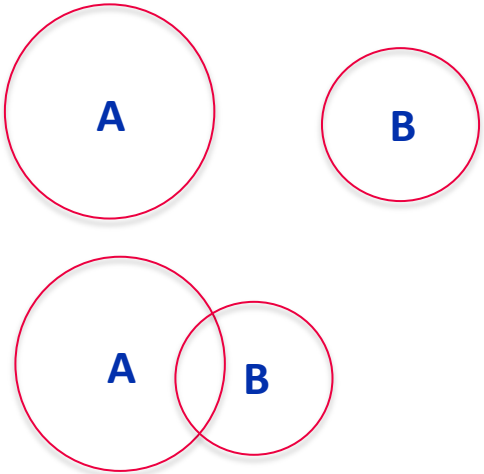
УНИВЕРСИТЕТ ИТМО

Types of random events: совместные и несовместные (compatible - incompatible)

Incompatible events $A_1 A_2 \ldots A_k = \varnothing$

Compatible events $A_1 + A_2 \ldots + A_k = I$

$A\overline{A} = \varnothing$ $A + \overline{A} = I$



Probability of events

УНИВЕРСИТЕТ ИТМО

Addition of probabilities for incompatible events:

$$P(\overline{A} \cup A) = P(\overline{A}) + P(A) = 1$$

$$P(\overline{A}) = 1 - P(A)$$

Addition of probabilities for compatible events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Conditional (условная) probability $P(B \mid A)$ **and multiplication of probabilities:**

$$P(A \cap B) = P(A)P(B \mid A)$$

Probabilistic independence (вероятностная независимость):

$$P(A \cap B) = P(A)P(B)$$

ITMOre than a UNIVERSITY

Random variables

Random variable

УНИВЕРСИТЕТ ИТМО

Discrete random variable **– a variable that can take on only specific values or countable set of values**

Probability law **– it is relation between certain value of random variable and it's probability**

$$p_k = P(\xi = \alpha_k), k = 1,2,...$$

Examples of discrete probability distributions:

- **Binomial probability distribution**
- **Hypergeometric probability distribution**
- **Multinomial probability distribution**
- **Negative binomial distribution**
- **Poisson probability distribution**

ITMO more than a UNIVERSITY

Random variable

УНИВЕРСИТЕТ ИТМО

Continuous random variable **- a variable that can take on any value in a specified range**

Probability distribution function (cumulative) **– probability that random outcome ξ takes on a value less than x**

$$F_\xi(x) = P(\xi < x),\ x \in R*$$

* Strictly speaking $F_\xi(x)$ is defined for discrete random variables, BUT $\xi, x \in N$

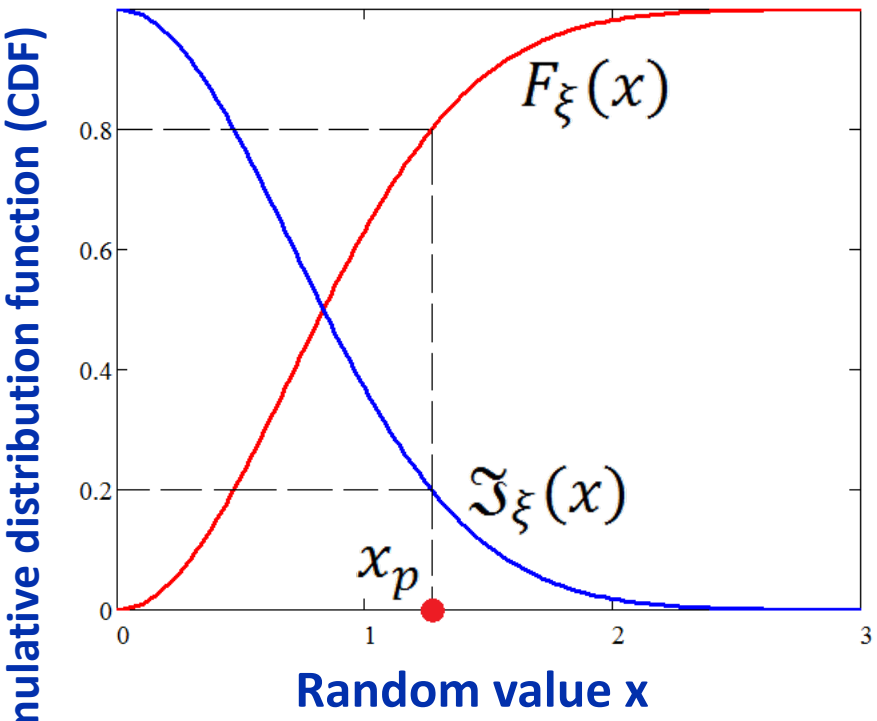
Distribution function properties (main):

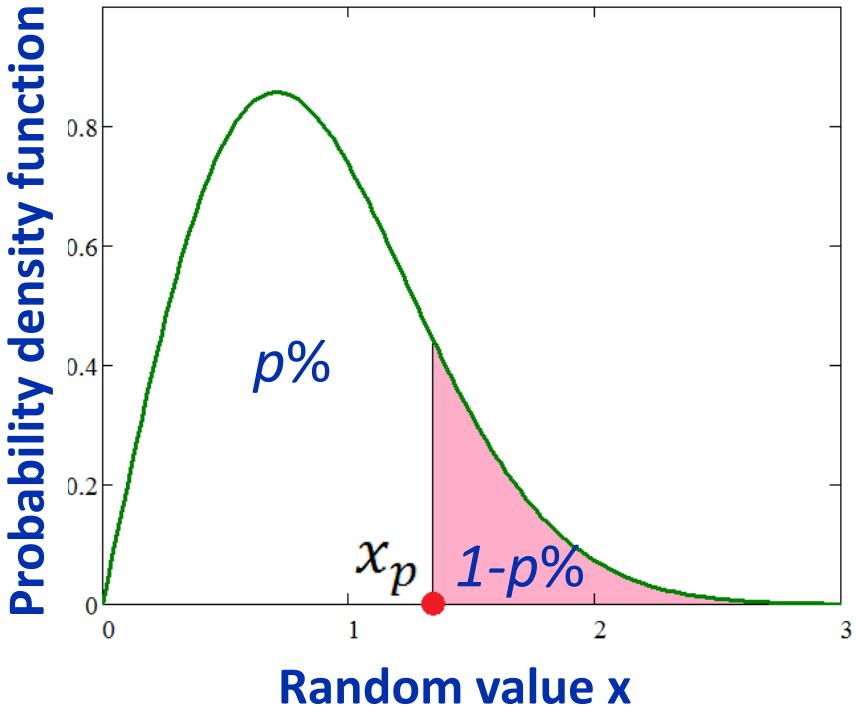
1. $0 < F_\xi(x) \le 1$
2. $P(a \le \xi < b) = F(b) - F(a)$

ITMO more than a UNIVERSITY

What is relation
between *probability distribution*
***function* and *probability density function*?**

CDF and PDF





p%-quantile :

$$x_p: \quad F_\xi(x_p) = p$$

Distribution density:

$$f_\xi(x) = \frac{dF_\xi(x)}{dx}$$

Random variable

УНИВЕРСИТЕТ ИТМО

What is relation between *probability distribution function* and *survival function*?

ITMO more than a UNIVERSITY

Useful model: categorical distribution

УНИВЕРСИТЕТ ИТМО

A categorical (multi-nomial) distribution is a generalized probability distribution of categories (k > 2) or discrete random value.

$$f(x = i \mid \boldsymbol{p}) = p_i, \quad \sum_{i=1}^{k} p_i = 1.$$

If there are only two categories k, the multi-nomial distribution - > Bernoulli distribution:

$$f(k; p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0. \end{cases}$$

Why it is so important?

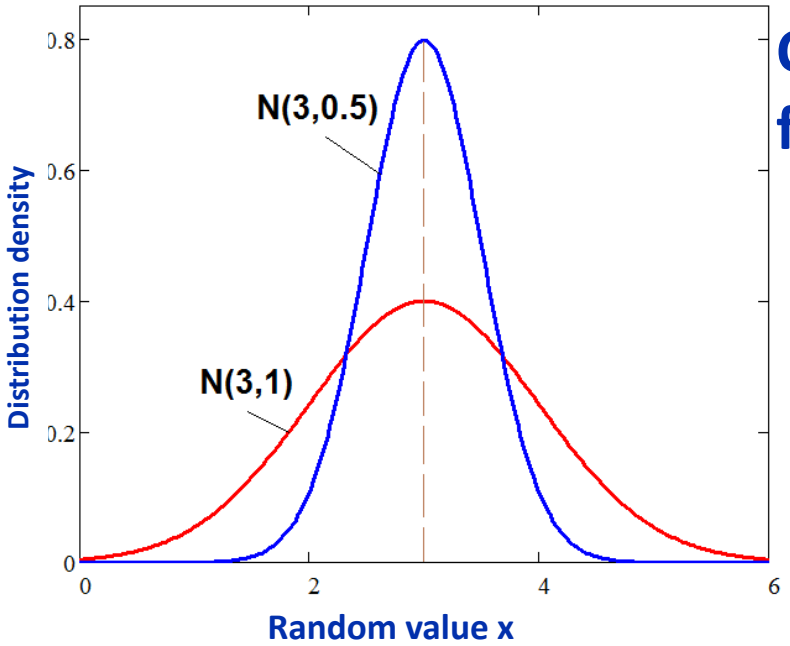
Plays an important role in the topic modeling in ML and NLP.

Useful model: Gaussian distribution

УНИВЕРСИТЕТ ИТМО

Distribution density: $$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma_\xi} exp\left[-\frac{(x - M_\xi)^2}{2\sigma_\xi^2}\right]$$

Nomenaclature: $\boldsymbol{N}(M_\xi, \sigma_\xi)$



Calculation of p%-quantile for Gaussian distribution:

$$x_p = M_\xi + \sigma_\xi u_p$$

Why it is so important?

The mixture of Gaussian distributions is a universal approximation for any smooth density function.

Random variable

What is the probability that random outcome *continuous* RV will have a certain value?

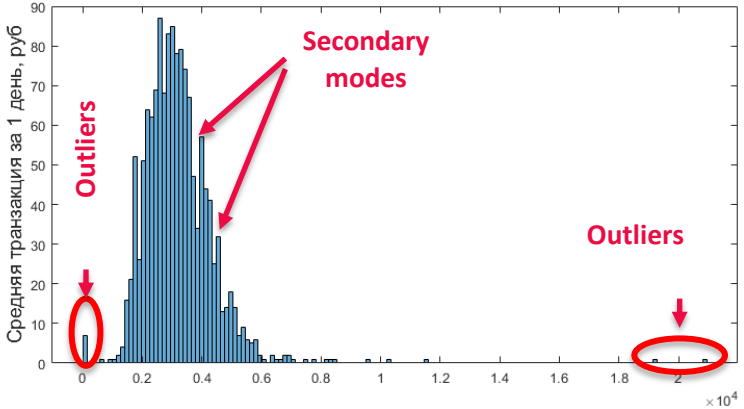
$$P(\xi = b) = ???$$

Nonparametric distribution estimators (1/2)

УНИВЕРСИТЕТ ИТМО

Histogram as way to display frequency characteristics of random values:

$$f^*(x) = \frac{n}{\Delta x n}, x - \frac{\Delta x}{2} < x_i < x + \frac{\Delta x}{2}$$

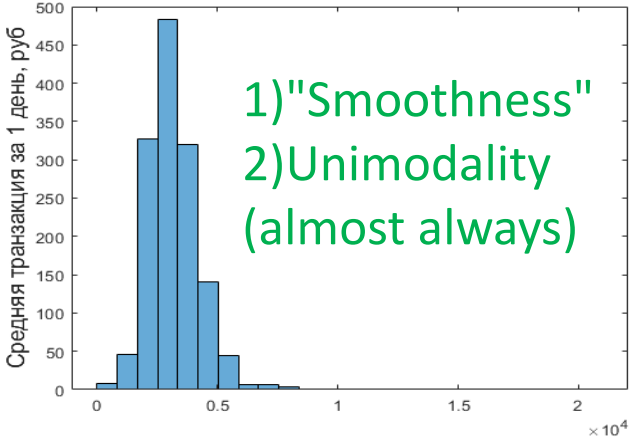


Assessment of number (m) of histogram columns:

1) $m = 1 + 3.32 \lg n$

2) $m = 5 \lg n$

3) **Own choice**



ITMOre than a UNIVERSITY

Nonparamteric distribution estimators (2/2)

УНИВЕРСИТЕТ ИТМО

The kernel smoothing method is often applied for "smooth" estimates of distribution density.

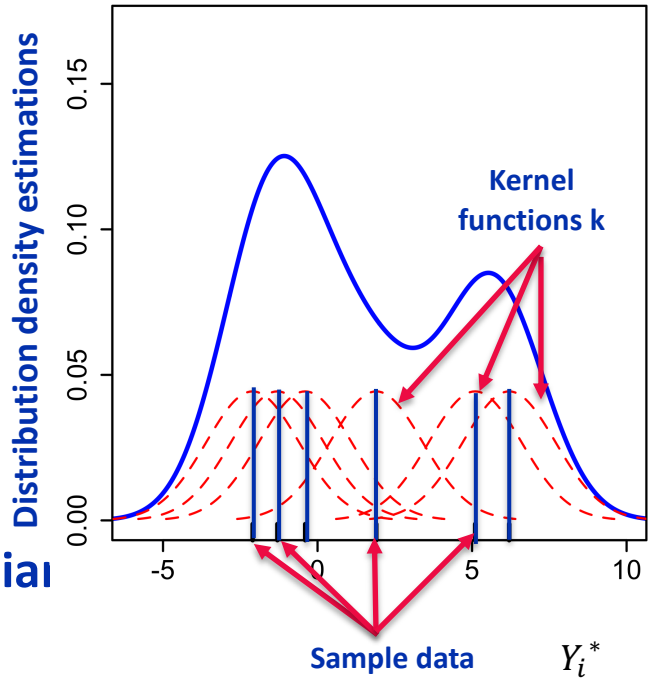
Essence: replacement of the rectangular not crossed histogram columns with the sum of curves (functions) with the centers in sample data:

$$f_{KDE}(y) = \frac{1}{Nh}\sum_i K\left(\frac{y - Y_i^*}{h}\right)$$

Sample data

Kenrel function

Smoothing parameter



There is a lot of kernel functions, but usually Gaussian function is enough:

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma_\xi} exp\left[-\frac{(x - M_\xi)^2}{2\sigma_\xi^2}\right]$$

Width of a window (h) is selected empirically, but it is possible to estimate on a formula: $h = \sigma^* N^{-0.2}$

Estimation of probability model

УНИВЕРСИТЕТ ИТМО

Probabilistic definition

1. **Probability distribution function (cumulative) for continuous variables:**

$$F_\xi(x) = P(\xi < x),\ x \in R$$

2. **Survival probability distribution function for continuous variables:**

$$\widetilde{F}_\xi(x) = P(\xi \geq x) = 1 - F_\xi(x)$$

Statistical definition

1. **Sample estimation of CDF:**

$$F^*(x) = \begin{cases} 0, & x \leq x_1 \\ \dfrac{k}{n}, & x_k < x \leq x_{k+1} \\ 1, & x > x_n \end{cases}$$

2. **Sample estimation of Survival CDF:**

$$1\text{-}F^*(x) = \begin{cases} 1, & x \leq x_1 \\ 1 - \dfrac{k}{n}, & x_k < x \leq x_{k+1} \\ 0, & x > x_n \end{cases}$$

ITMO more than a UNIVERSITY

Estimation of probability model

УНИВЕРСИТЕТ ИТМО

Probabilistic definition

3. Probability (density) function:

$$f_\xi(x) = F'_\xi(x)$$

Probability density function properties (main):

1. $\int_{-\infty}^{\infty} f(x)dx = 1$
2. $P(a \le \xi < b) = \int_a^b f(x)dx$

ITMO more than a UNIVERSITY

Statistical definition

3. Sample estimation of PDF:

$$f^*(x) = \frac{n(x)}{\Delta x n}, x - \frac{\Delta x}{2} < x_i < x + \frac{\Delta x}{2}$$

Kernel density estimation is another way for PDF estimation:

$$fk(y) := \frac{1}{N \cdot h} \cdot \sum_i K\left(\frac{y - A_i}{h}\right)$$

$$K(x) := \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left(\frac{-x^2}{2}\right)$$

Random variable

УНИВЕРСИТЕТ ИТМО

What are *moments* of probability distribution?

What is *mathematical expectation* from probability theory point of view? How it could be estimated in statistics?

ITMOre than a UNIVERSITY

Estimation of probability model

УНИВЕРСИТЕТ ИТМО

Probabilistic definition

4. High-order moments of probability distribution

$$m_k = \int_R (x - c)^k f_\xi(x) dx$$

If $c = 0$ → m_k is a **raw** moment

If $c = M[X]$ → m_k is a **central** moment

First raw moment – *mathematical expectation*

$$M[x] = m_1 = \int_R x f_\xi(x) dx$$

Statistical definition

4. High-order moments estimation

First raw moment point unbiased estimation – *mean value*

$$\bar{x} = \frac{\sum_n x}{n}$$

ITMOre than a UNIVERSITY

Estimation of probability model

УНИВЕРСИТЕТ ИТМО

Probabilistic definition

4. **High-order moments of probability distribution**

Second central moment – *variance*

$$V[x] = \int_{-\infty}^{\infty} (x - M[x])^2 f_{\xi}(x) dx$$

Statistical definition

4. **High-order moments estimation**

Second central moment point *biased* estimation

$$V^* = \frac{\sum_n (x - \bar{x})^2}{n}$$

Second central moment point *unbiased* estimation

$$V^{**} = \frac{\sum_n (x - \bar{x})^2}{n - 1}$$

For big samples *V*=V***

ITMO more than a UNIVERSITY

Estimation of probability model

УНИВЕРСИТЕТ ИТМО

What is *quantile (percentile)*
of probability distribution?
Median?
Mode?

ITMOre than a UNIVERSITY

Estimation of probability model

УНИВЕРСИТЕТ ИТМО

Probabilistic definition

5. Quintiles (percentiles) of probability function:

$$x_p: \quad F_\xi(x) = p$$

$x_{0,25}, x_{0,75}$ - Quartiles

$Me = x_{0,5}$ - Median

$Q = x_{0,75} - x_{0,25}$ - Inter quartile range

6. Mode of probability function:

$$x_m: \quad f_\xi(x_m) \to \max$$

Statistical definition

5. Quintiles (percentiles) sample estimation using *order statistics*

Sample in ascending order:

$$x_1 < x_2 < x_i < \ldots < x_n$$

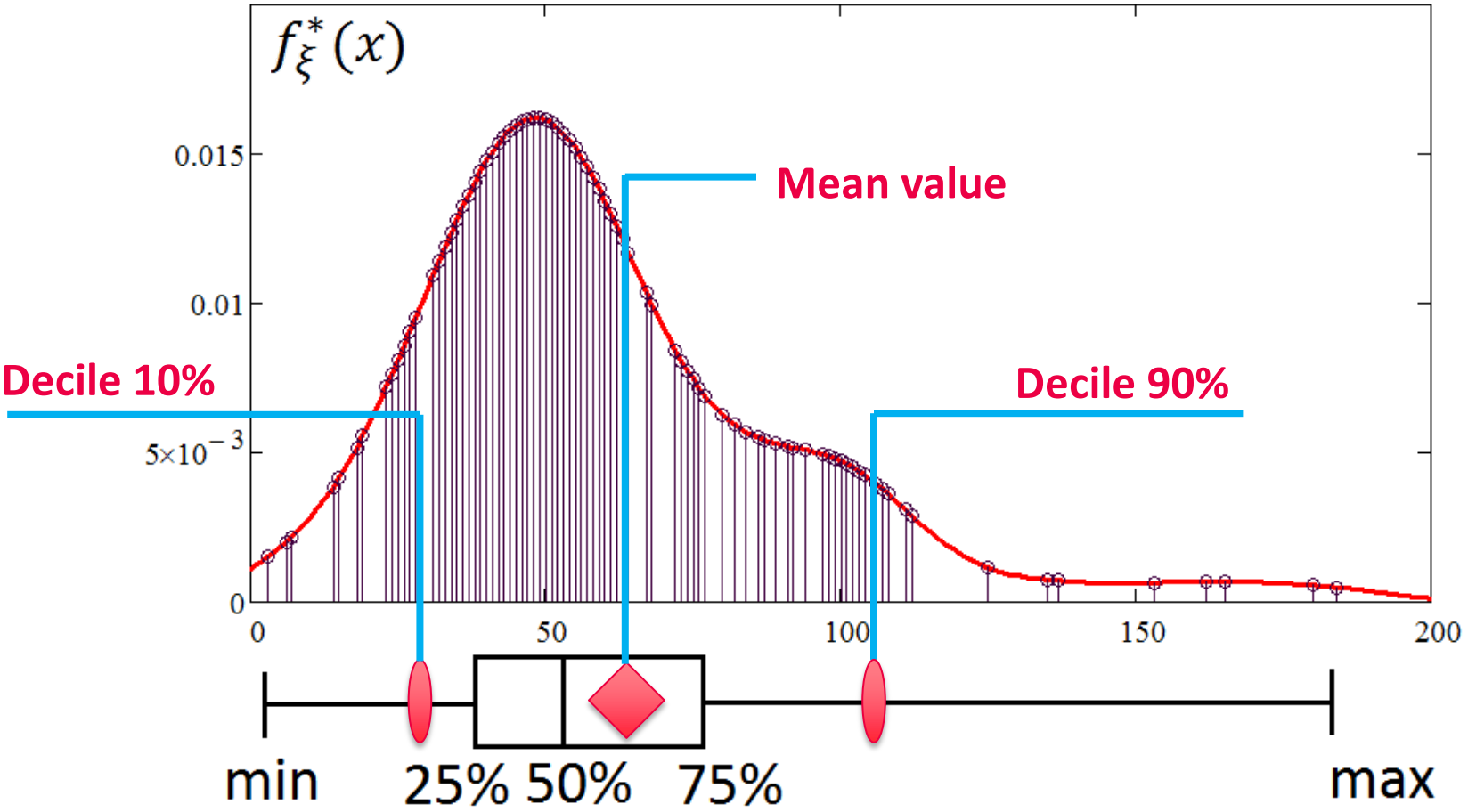
$$x_p = x_i, i = np + 1$$

6. Mode of probability function:

$$Mode^* = \max(f^*(x))$$

ITMO more than a UNIVERSITY

Mapping of quantiles: boxes with moustaches



Is sample mean value always enough? Why we need mode and median?

Measures of central tendency

УНИВЕРСИТЕТ ИТМО

Is sample mean value always enough?

Sample mean value: $M_\xi^* \stackrel{\text{def}}{=} \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

(is suitable only for real numbers)

For example: average clients per day ~~=~~ **10,42**

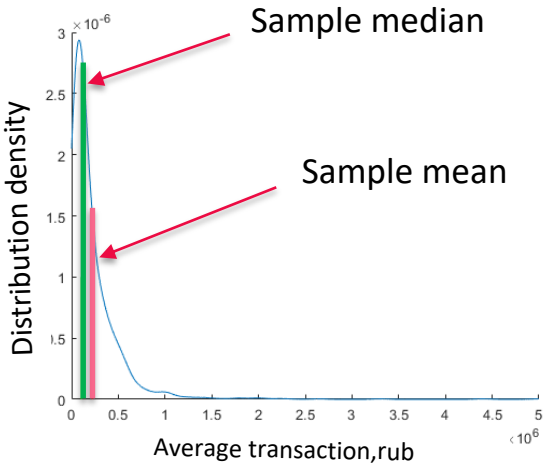
Sample median: $Me = x_{p=0.5}$

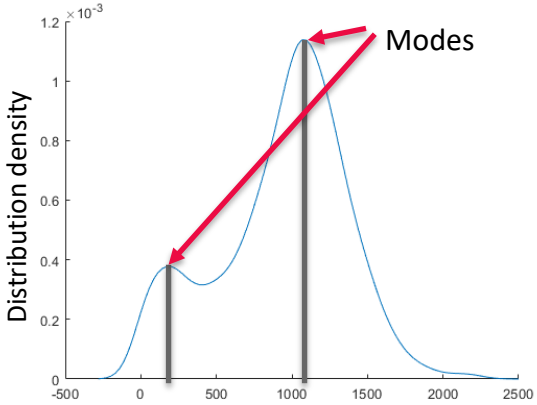
(is suitable for integers, real and ordinal numbers)

Sample mode: $Md = argmax(f^*)$

(is suitable for all numbers)

For all scales, except continuous, the mode is defined by the value having bigger repeatability than the next values.





Measures of deviation

УНИВЕРСИТЕТ ИТМО

The measure of spread characterizes "density" of data around the central trend

Sample variance:

$$D_\xi^* \stackrel{\text{def}}{=} s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Sample root-mean-square deviation :

$$s = \left(D_\xi^*\right)^{1/2}$$

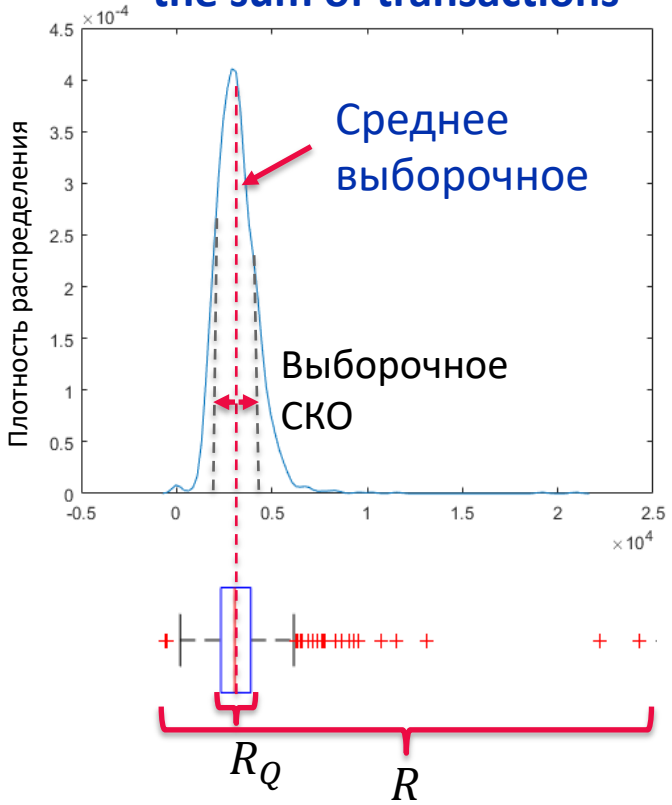
Sample range:

$$R = X_{max} - X_{min}$$

Interquartile distance:

$$R_Q = X_{0.75} - X_{0.25}$$

Example: Estimate of outliers for the sum of transactions

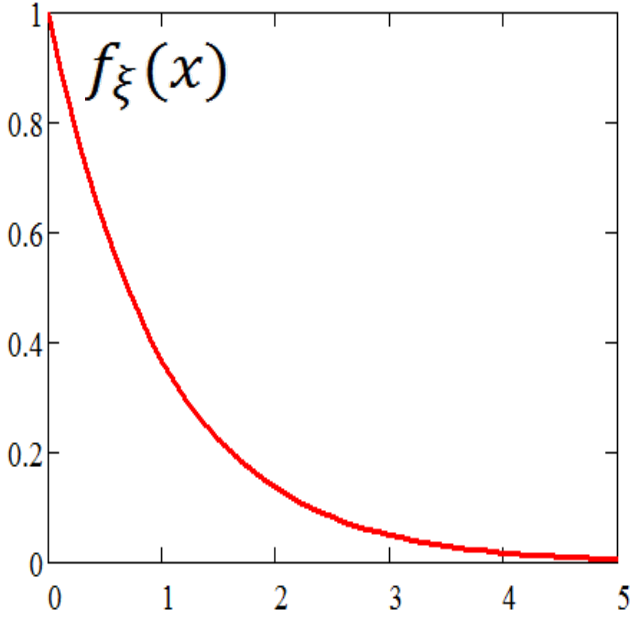


What brilliant feature has ordered statistics (quantiles) vs. parametric estimation of distributions?

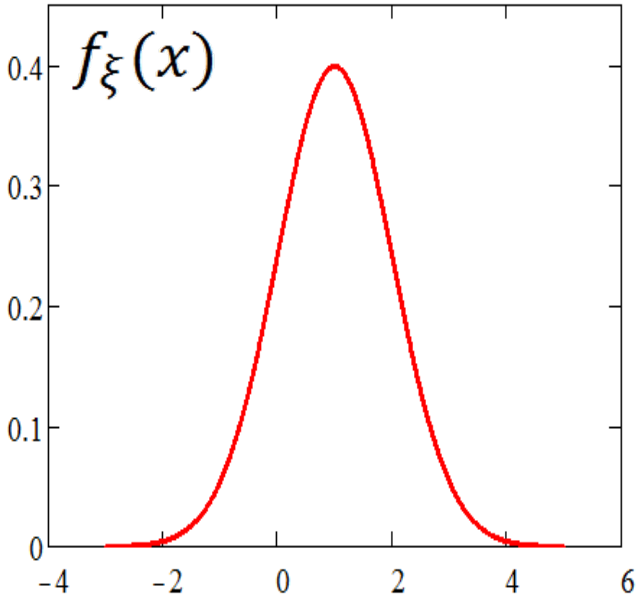
Variety of distributions

УНИВЕРСИТЕТ ИТМО

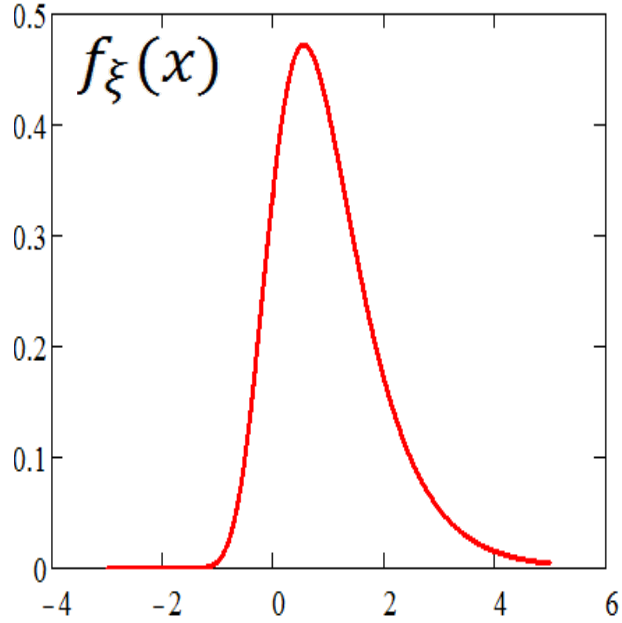
Exponential distribution



Gaussian distribution



Gumbel distribution



ITMO more than a UNIVERSITY

Distribution characteristics: $\bar{x} = 1, \quad s^2 = 1$

Characteristic quantiles and their complexes

УНИВЕРСИТЕТ ИТМО

Selective median:

$$Me^* = \begin{cases} x_{(N/2+1)}, & N - \text{odd} \\ (x_{(N/2+1)} + x_{(N/2)})/2, & N - \text{even} \end{cases}$$

Interquartile distance :

$$Q = x_{(\lfloor 3N/4 \rfloor + 1)} - x_{(\lfloor N/4 \rfloor + 1)}$$

Interdecile distance:

$$D = x_{(\lfloor 9N/10 \rfloor + 1)} - x_{(\lfloor N/10 \rfloor + 1)}$$

Sample range:

$$R = x_{(N)} - x_{(1)}$$

k-average amplified range:

$$R_{(k)} = \frac{1}{k}\left(\sum_{j=0}^{k-1} x_{(N-j)} - \sum_{j=1}^{k} x_{(j)}\right)$$

Interval estimates

УНИВЕРСИТЕТ ИТМО

What *is interval estimate*?

In what cases we use *interval estimates?*

What is difference between *probability interval* and *confidence interval?* (3 points)

ITMO more than a UNIVERSITY

Interval estimates

Probability interval should not be confused with *Confidence interval*

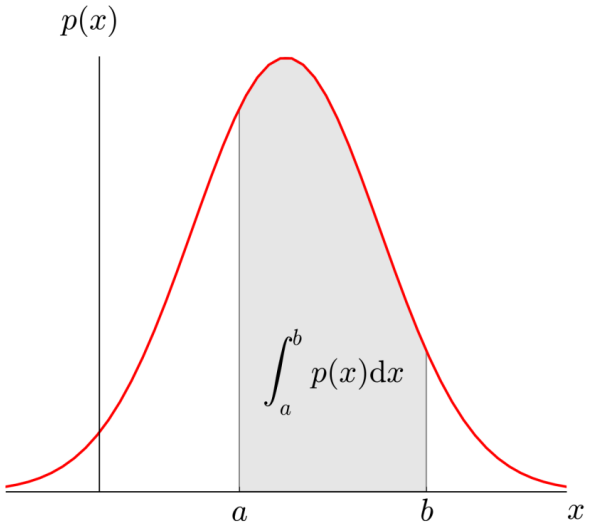
More general definition:

Probability interval - is an estimate of an interval in which random outcome of random variable will fall, with a certain probability, given by known or observed probability distribution function

Probability interval definition is followed from properties of probability distribution functions:

$$P(a \le \xi < b) = F(b) - F(a)$$

$$P(a \le \xi < b) = \int_a^b f(x)dx$$



Interval estimates

УНИВЕРСИТЕТ ИТМО

Confidence interval is a particular case for probability interval for distribution function parameter estimates:

- Confidence interval is based on known distribution function for estimate of parameter.
- Confidence interval range can vary from sample to sample due to its size.
- Confidence interval can be built using known and unknown parameters

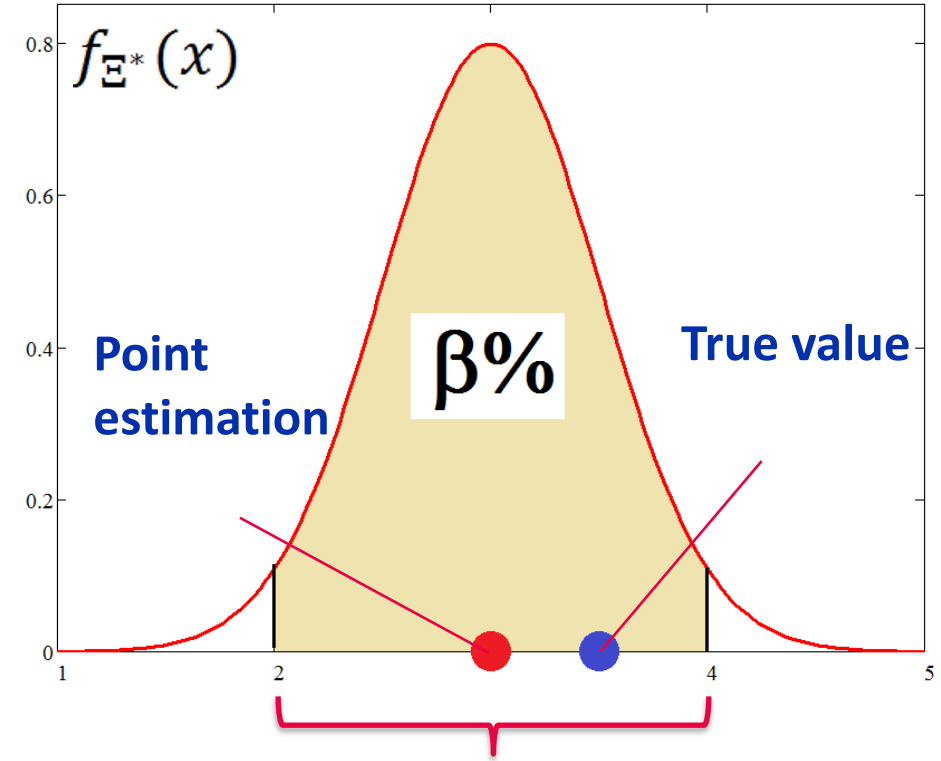
$$P[\widetilde{\Theta} - u_{1-\alpha/2}\sigma_{\widetilde{\Theta}} < \Theta < \widetilde{\Theta} + u_{1-\alpha/2}\sigma_{\widetilde{\Theta}}] \approx 1 - \alpha$$

where α is significance level, (1-α) is a confidence level.

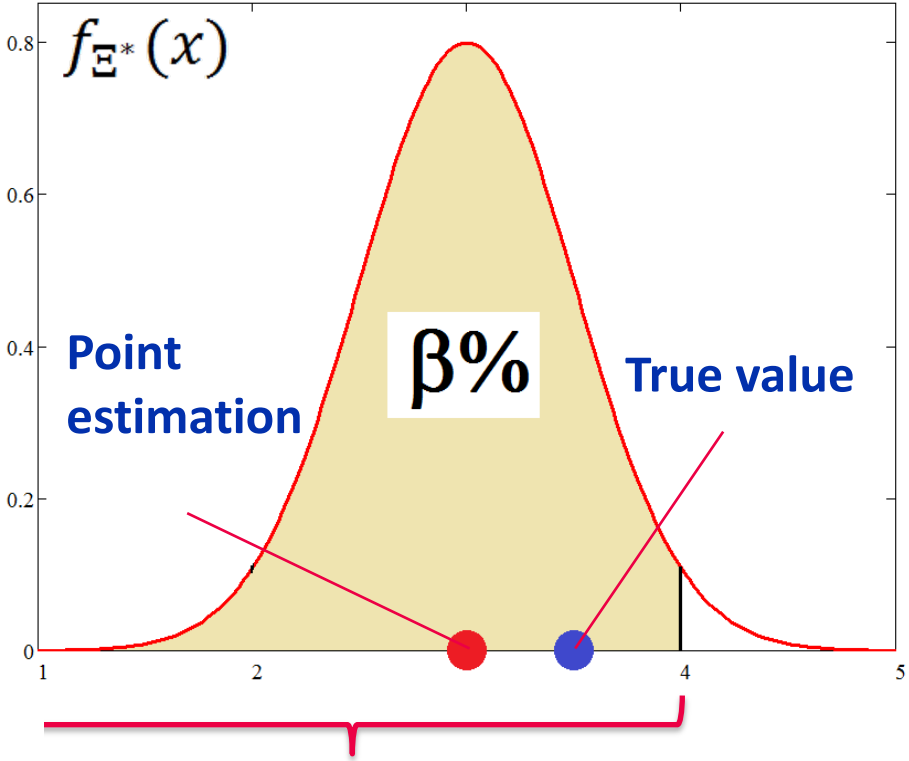
ITMO more than a UNIVERSITY

The higher the confidence level, the wider bounds of confidence interval.

Interval estimates



β% *two-sided confidence interval*



β% one-sided confidence interval

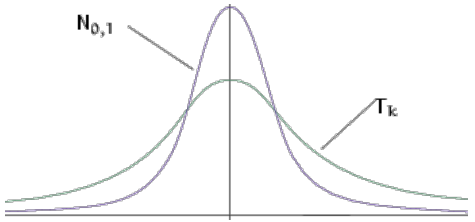
Interval estimates

УНИВЕРСИТЕТ ИТМО

Confidence interval for mean (σ is a sample estimate):

$$P[\mu^* - t_{1-\alpha/2}\frac{\sigma_\mu}{\sqrt{n}} < \mu < \mu^* + t_{1-\alpha/2}\frac{\sigma_\mu}{\sqrt{n}}] \approx 1-\alpha$$

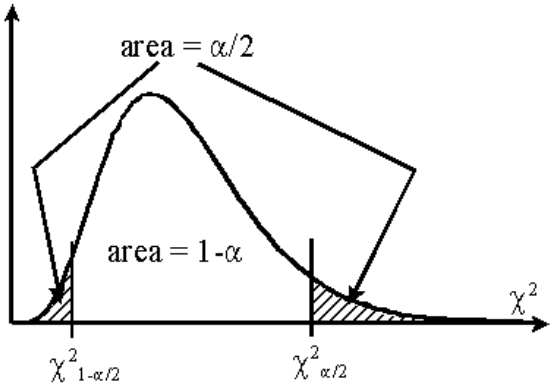
t -Student distribution



Confidence interval for variance:

$$P[\frac{D^*(n-1)}{\chi^2_{\alpha/2}} < D < \frac{D^*(n-1)}{\chi^2_{1-\alpha/2}}] \approx 1-\alpha$$

χ^2 probability density:



Confidence interval for standard deviation:

$$P[\frac{\sigma^*\sqrt{n-1}}{\chi_{\alpha/2}} < \sigma < \frac{\sigma^*\sqrt{n-1}}{\chi_{1+(1-\alpha)/2}}] \approx 1-\alpha$$

ITMOre than a UNIVERSITY

Interval estimates

УНИВЕРСИТЕТ ИТМО

Prediction interval **- is an estimate of an interval in which future observations will fall, with a certain probability, given what has already been observed**

From "Wikipedia"

Prediction interval = Probability interval

In other word *prediction interval* shows that you expect next predicted value will be within boundaries of calculated distribution of observed population.

Prediction interval **tells you about the distribution of values, and about its uncertainty.**

ITMO more than a UNIVERSITY

Interval estimates

УНИВЕРСИТЕТ ИТМО

What is tolerance interval? (3 points)

ITMO more than a UNIVERSITY

Tolerance interval

УНИВЕРСИТЕТ ИТМО

Tolerance interval **is an interval within which, with some confidence level, a specified proportion of a sampled population falls.**

Tolerance interval may be described as confidence interval for estimate of probability interval.

$$P\left[\int_{x_{(a)}}^{x_{(b)}} f(x)dx \geq \gamma\right] = \beta$$

$$P\left[F(x_{(b)}) - F(x_{(a)}) \geq \gamma\right] = \beta$$

where $x_{(a)}$ - quintiles of distribution, γ - width of probability interval,

β - confidence level

ITMO more than a UNIVERSITY

Estimation of distribution parameters

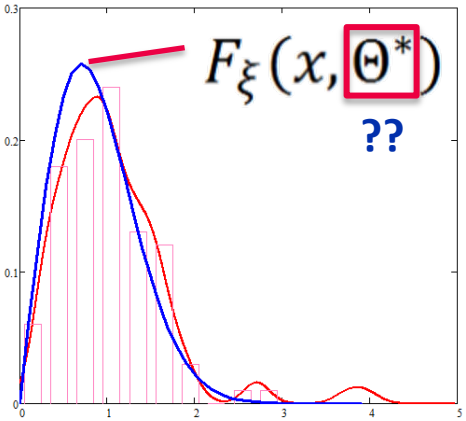
УНИВЕРСИТЕТ ИТМО

Purpose: to understand under what law the analysed random value lives (the income of the client, turnover of the bank branch, etc.)

Means: parametrical distributions

$$F_\xi^*(x) = F_\xi(x, \Theta^*)$$

Essence of estimation of distribution parameters: search of parameters which minimize a divergence betwee theoretical and empirical distributior



Known distribution

ITMOre than a UNIVERSITY

Estimation of distribution parameters

УНИВЕРСИТЕТ ИТМО

There are several methods of parameter estimation in statistics:

1. Method of moments
2. Method of quantiles (inverse distribution method)
3. Maximum likelihood method
4. Least square method

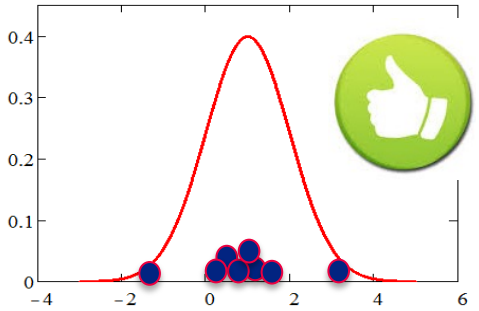
ITMO more than a UNIVERSITY

Maximum likelihood method

УНИВЕРСИТЕТ ИТМО

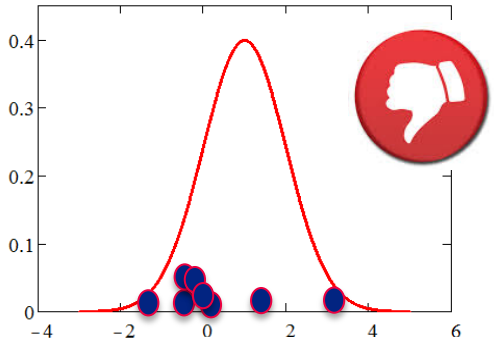
Likelyhood function:

$$L(x_1, \dots, x_N | \Theta) = \prod_{i=1}^{N} f_\xi(x_i, \Theta)$$



Idea:

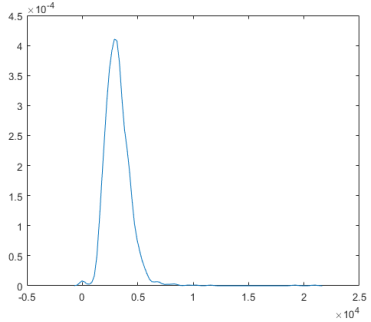
$$L(x_1, \dots, x_N | \Theta) \underset{\Theta}{\rightarrow} max$$

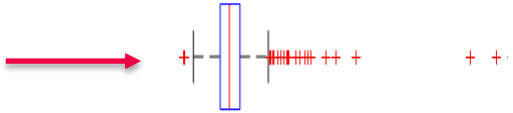


The transformed task:

$$\ln(L) \underset{\Theta}{\rightarrow} max$$

Example: distribution
of transaction sum
and estimation of likelyhood of
Gaussian distribution





Least Squares Method

УНИВЕРСИТЕТ ИТМО

Residuals as Euclidean distance:

$$d(x_p) = (x_p^* - x_p(\Theta))^2$$

The idea (squares sum minimization of residuals by L characteristic quantiles):

$$\sum_{k=1}^{L} (x_{p_k}^* - x_{p_k}(\Theta))^2 \underset{\Theta}{\rightarrow} min$$

Method implementation (equation system):

$$\sum_{k=1}^{L} (x_{p_k}^* - x_{p_k}(\Theta)) \frac{\partial x_{p_k}(\Theta)}{\partial \theta_j} = 0, \qquad j = 1, \dots, l$$

LS method for linaerly scalable variables

Quintile of linearly scalable variable:

$$x_{p_k}(\Theta) = \theta_1 + \theta_2 U_{p_k}$$

$$U_{p_k} = F^{-1}_{\xi(0)}(p_k)$$ - **quantile of normilized model distribution**

LS implementation:
System with 2 linear algebraic equations

$$\begin{cases} L\theta_1 + \theta_2 \sum_{k=1}^{L} U_{p_k} = \sum_{k=1}^{L} x^*_{p_k} \\ \theta_1 \sum_{k=1}^{L} U_{p_k} + \theta_2 \sum_{k=1}^{L} U^2_{p_k} = \sum_{k=1}^{L} U_{p_k} x^*_{p_k} \end{cases}$$

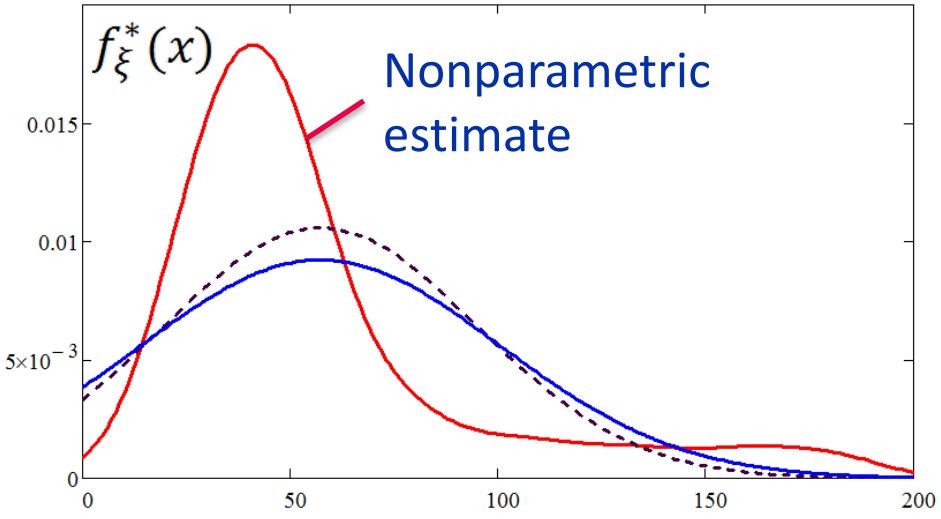
The simplest approximation: throughout all sample

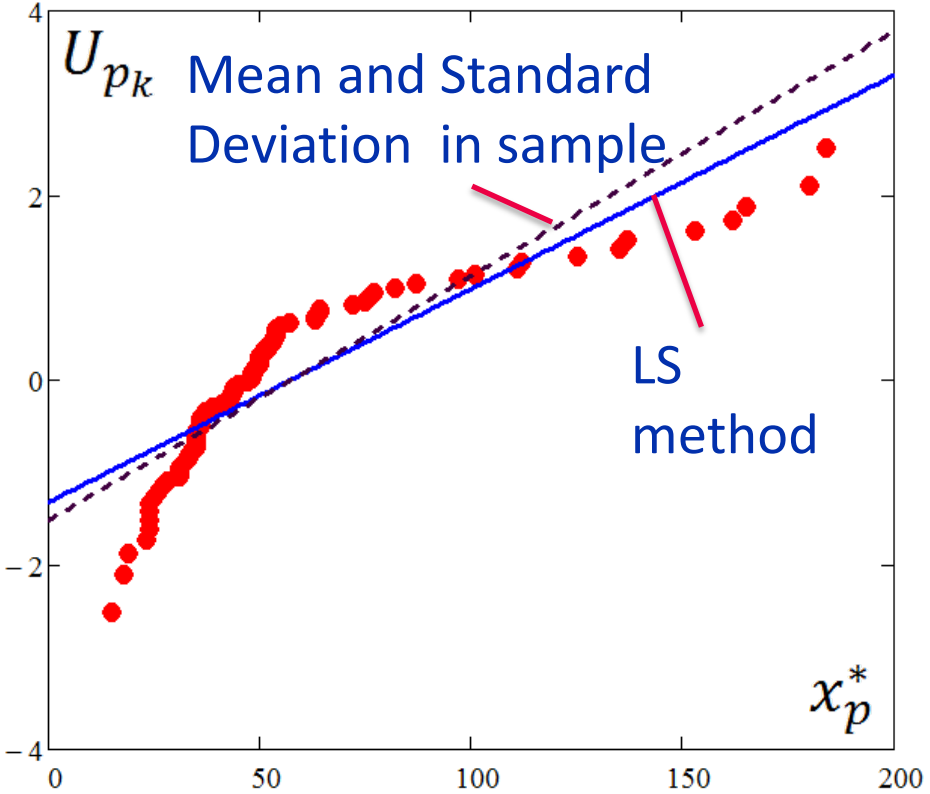
УНИВЕРСИТЕТ ИТМО

Task: approximate model distribution to sample, which not quite corresponds to it

All sample quantiles are eqvivalent

$$L = N$$





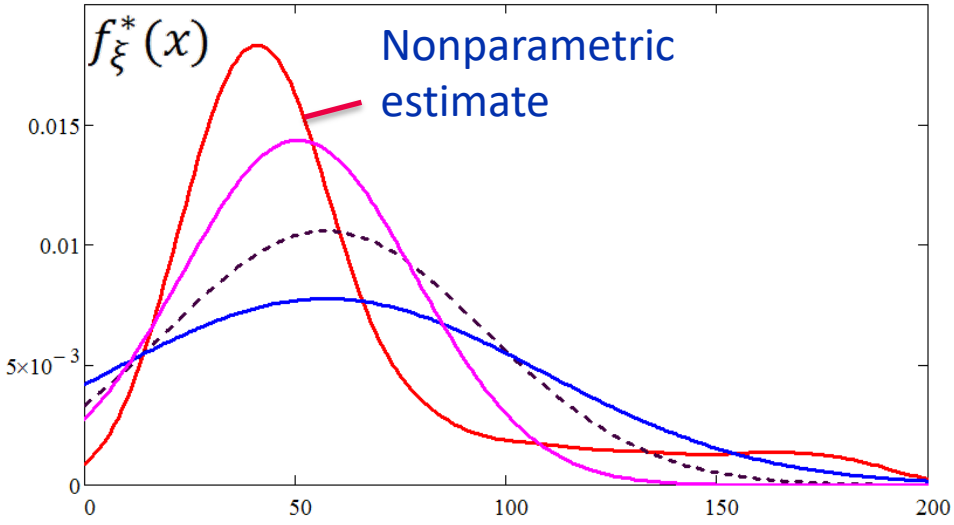
Weighted LS: consider aspects

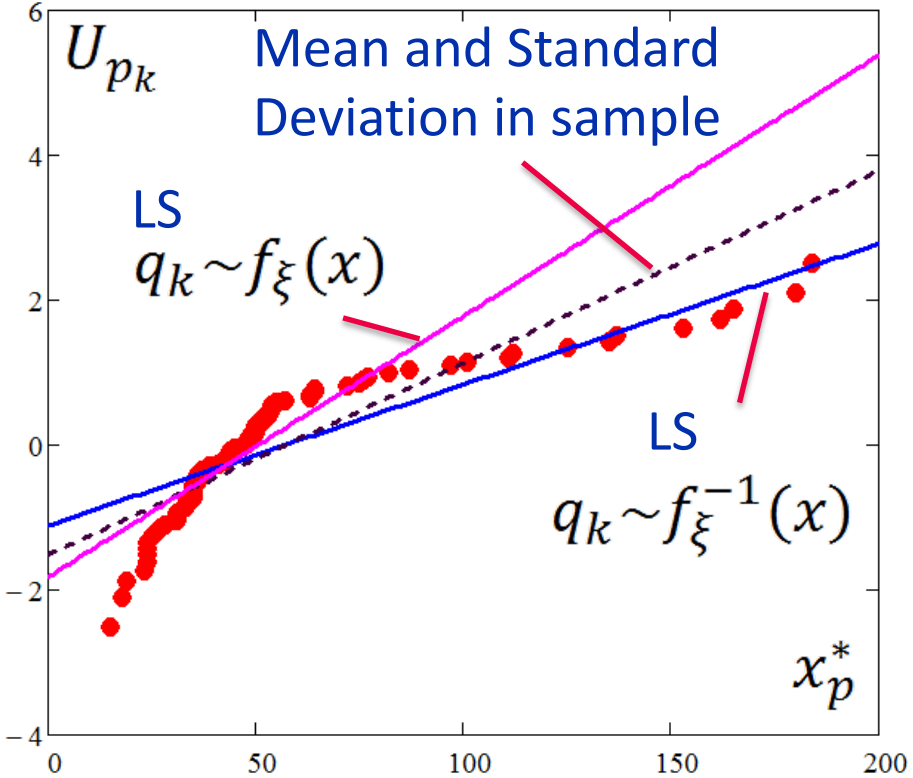
$$\sum_{k=1}^{L} q_k \left(x^*_{p_k} - x_{p_k}(\Theta)\right)^2 \underset{\Theta}{\rightarrow} min$$

Objective function with weights q_k

All sample quantiles are not eqvivalent

$$L = N$$





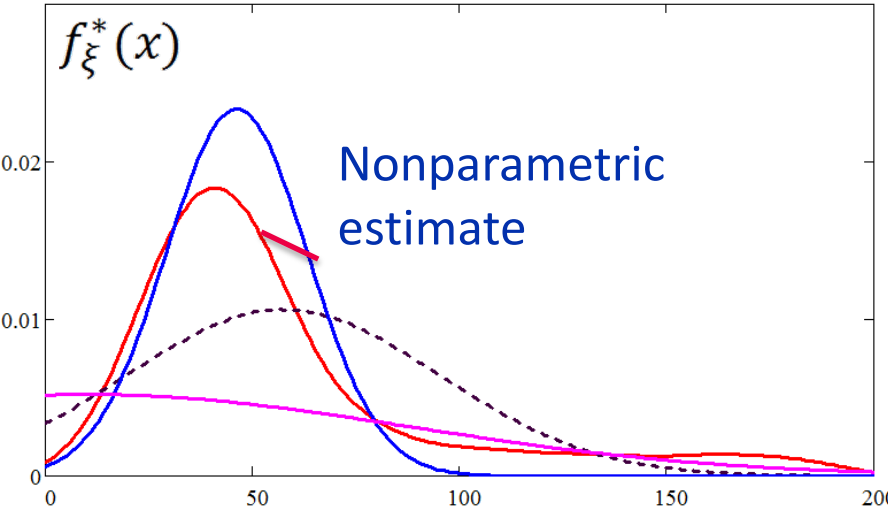
LS method with selected quantiles

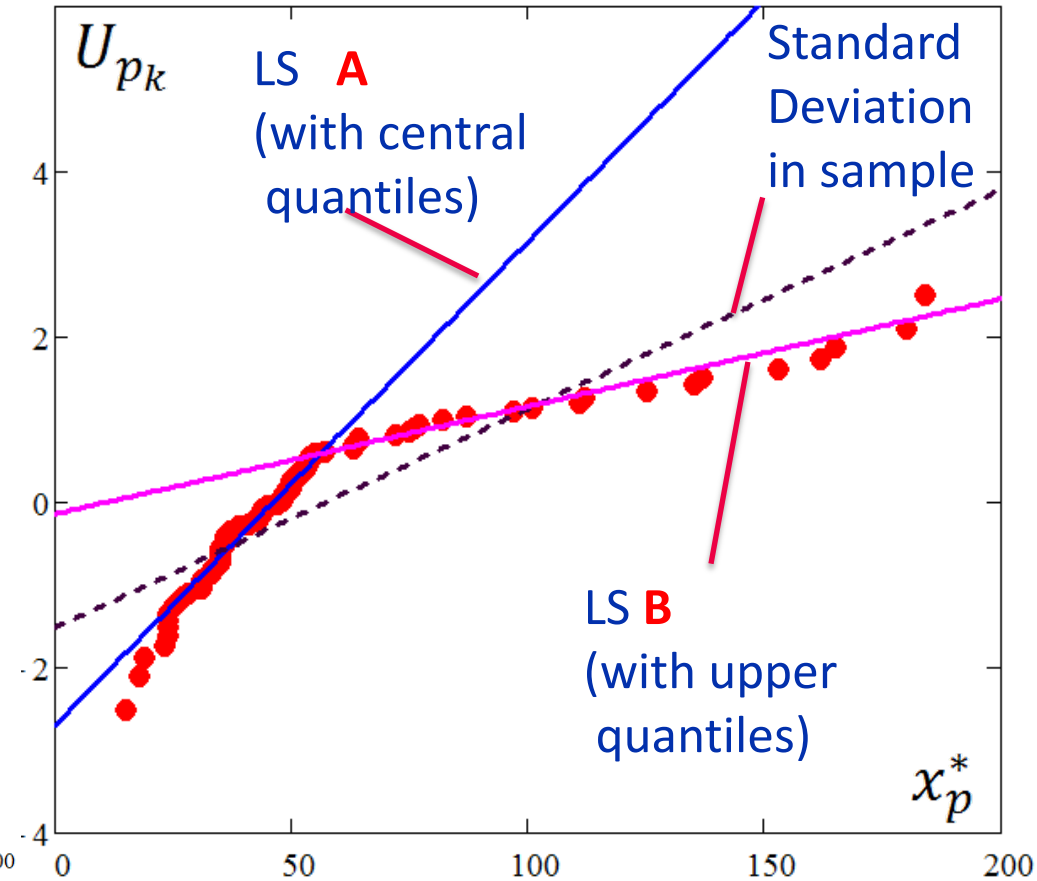
УНИВЕРСИТЕТ ИТМО

Presumtion: selected quantiles are eqvivalent $L < N$

A: quantiles (10%,25%,35%, 50%, 65%, 75%)

B: quantiles (75%,85%,90%,95%, 97%,99%)

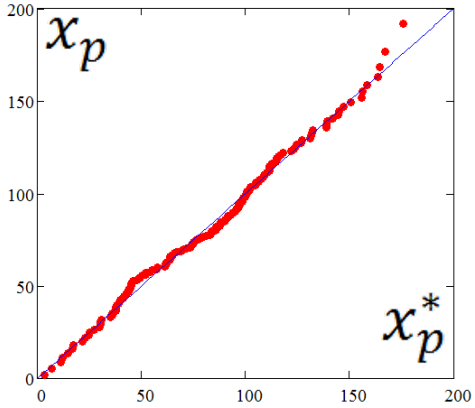


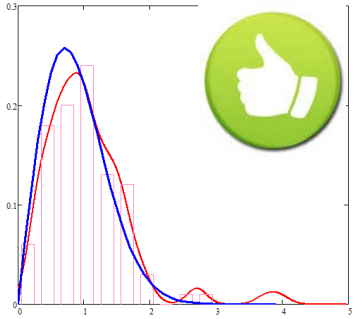


Quality analysis of the estimated parameters

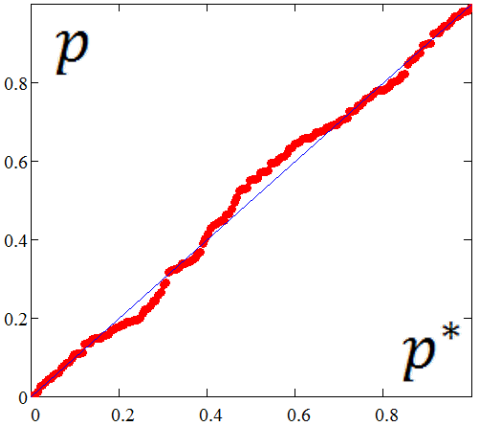
УНИВЕРСИТЕТ ИТМО

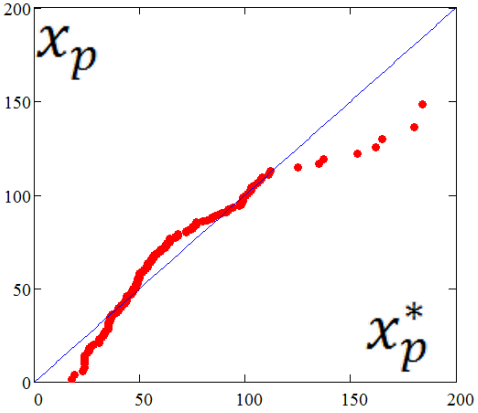
Quantile biplot (*q-q plot*)

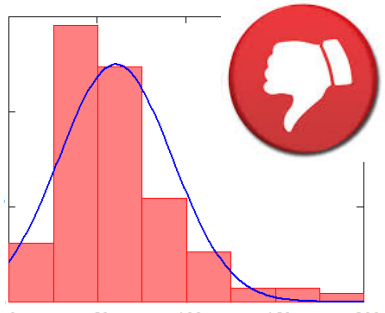


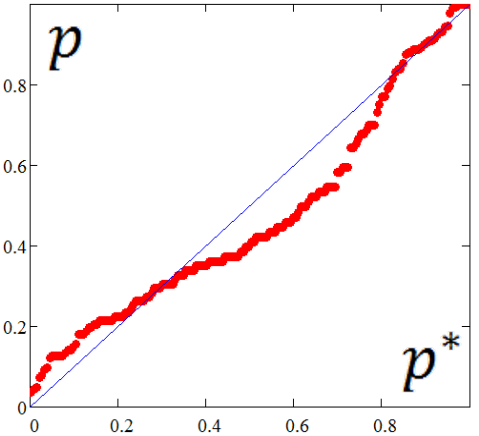


Probability biplot (*p-p plot*)









Goodness-of-fit criteria

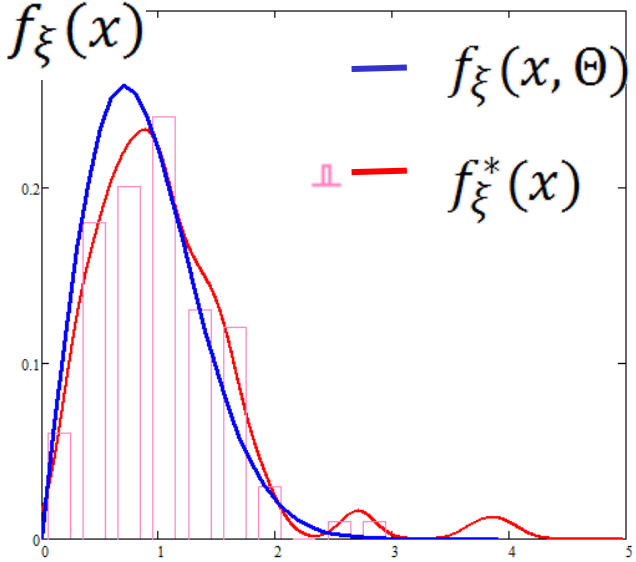
УНИВЕРСИТЕТ ИТМО

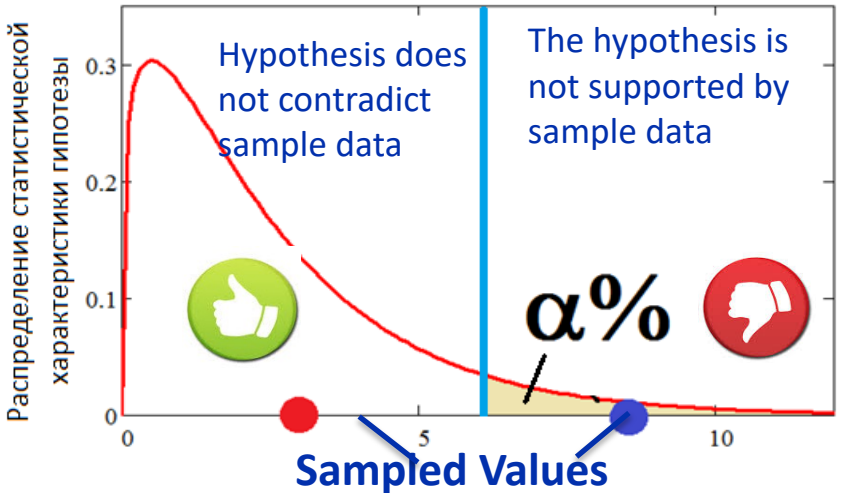
Goodness-of-fit criterion: a mechanism for testing the null hypothesis by comparing selected data with a theoretical “standard”

The idea: the characteristic of the residual is selected to compare reality and the “standard”, and a parametric distribution model is created for this characteristic.

Null hypothesis - H_0 **:** $f_\xi(x,\Theta) = f_\xi^*(x)$

Alternative H_1 **:** $f_\xi(x,\Theta) \neq f_\xi^*(x)$



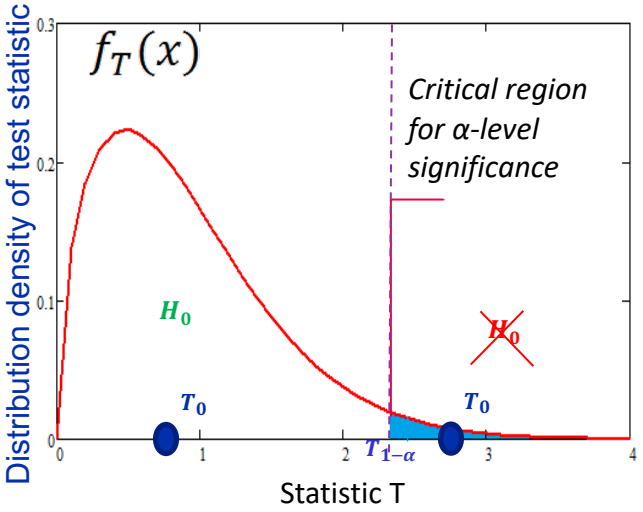


When we need to compare probability distributions?

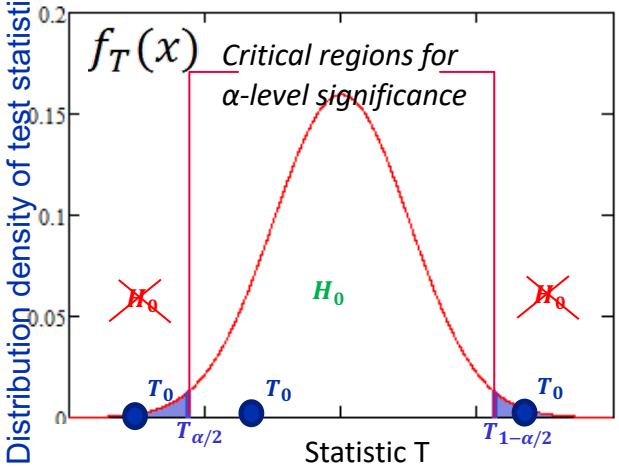
Tests: comparison of statistics

УНИВЕРСИТЕТ ИТМО

One-sided test



Both-sided test



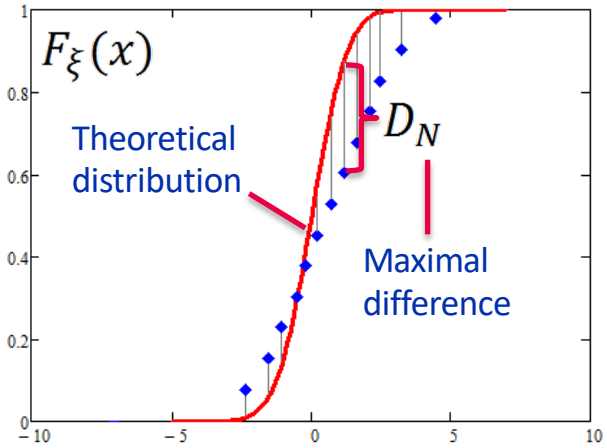
Error of non-parametric testing (*I kind error*) – possibility to reject correct hypothesis with α probability

ITMOre than a UNIVERSITY

Kolmogorov test

УНИВЕРСИТЕТ ИТМО

Goodness of fit testing



Estimated statistic:

$$D_N = max|F^*(x) - F(x)|$$

Test (*rejection of* H_0):

$$\sqrt{N} D_N \geq T_{1-\alpha}$$

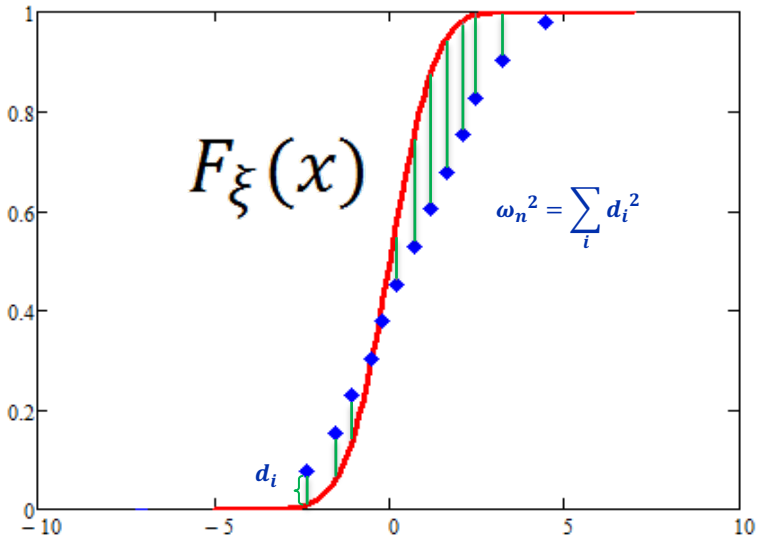
$T_{1-\alpha}$ **- quantile of Kolmogorov distribution**

ITMO more than a UNIVERSITY

Omega squared test (Cramér–von Mises)

УНИВЕРСИТЕТ ИТМО

Goodness of fit testing



Estimated statistic:

$$N\omega_N^2 = \frac{1}{12N} + \sum_{i=1}^{N}\left[F\left(x_{(i)}\right) - \frac{2i-1}{2N}\right]^2$$

Test (*rejection of* H_0):

$$N\omega_N^2 \geq T_{1-\alpha}$$

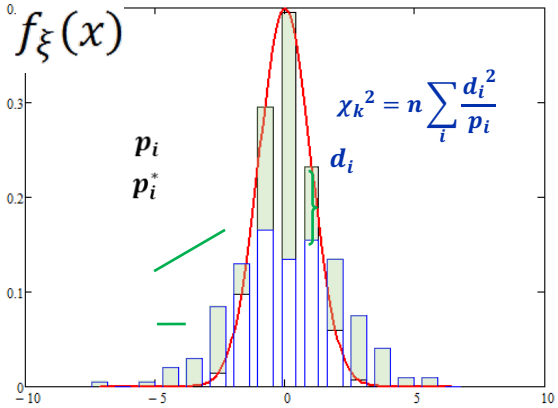
$T_{1-\alpha}$ **- quantile of omega distribution**

ITMO more than a UNIVERSITY

Pearson's test

УНИВЕРСИТЕТ ИТМО

Goodness of fit testing for grouped data



Random variable *x*

Estimated statistic:

$$\chi_N^2 = N\sum_{j=1}^{M}\frac{(p_j^* - p_j)^2}{p_j}$$

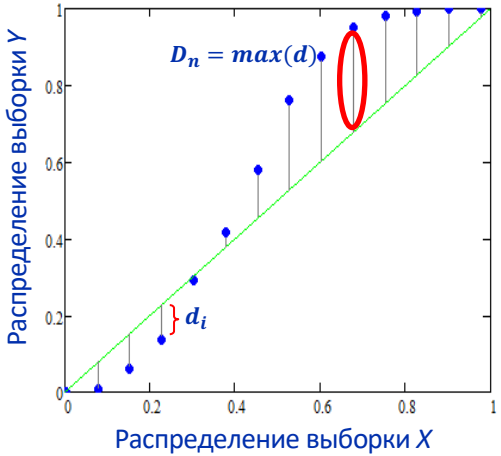
Test (*rejection of* H_0):

$$\chi_N^2 \geq T_{1-\alpha}$$

$T_{1-\alpha}$ **- quantile of** χ_{N-1}^2 **distribution**

ITMO more than a UNIVERSITY

Homogeneity tests



Null hypothesis $H_0: f_\xi^*(x) = g_\xi^*(x)$

Kolmogorov test ⟹ **Smirnov**

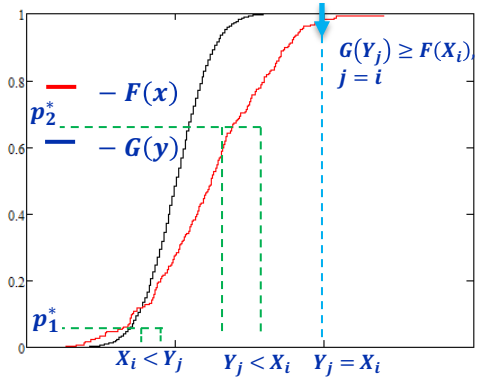
$$D_{NM} = \max|F^*(x) - G^*(x)|,$$

Test ω^2 ⟹ **Rosenblatt**

$$\frac{NM}{N+M}\omega_{NM}^2 = \frac{1}{N+M}\left(\frac{1}{6} + \frac{1}{M}\sum_{i=1}^{N}(R_i - i)^2 + \frac{1}{N}\sum_{j=1}^{M}(S_j - j)^2\right) - \frac{2}{3}$$

Rules for testing the same with goodness of fit criterions

Wilcoxon rank-sum test



Null hypothesis H_0: $F(x) = G(x)$

Alternative hypothesis H_2: $F(x) < G(x)$

Samples $(x_1, \dots, x_N)$ **and** $(y_1, \dots, y_M)$

Estimated statistic:

$$U_{st} = \frac{\min\{U_X, U_Y\} - \frac{1}{2}MN}{\sqrt{\frac{1}{12}MN(M+N+1)}}$$

$$U_Y = \sum_{i=1}^{N}\sum_{j=1}^{M}[x_i > y_j] \qquad U_X = \sum_{i=1}^{N}\sum_{j=1}^{M}[x_i < y_j]$$

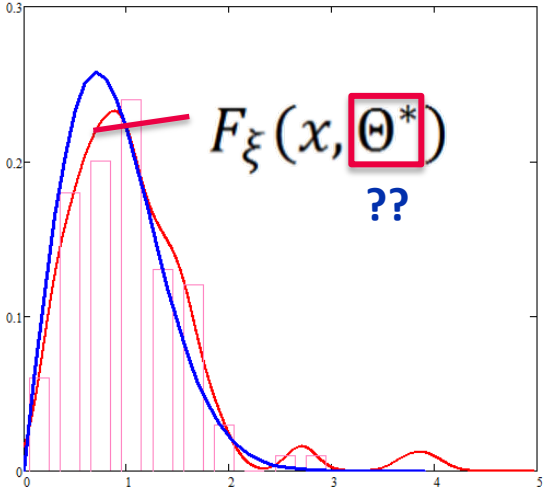
Test (*rejection of* H_0):

H_0

$$U_{st} \notin [Z_{\alpha/2}, Z_{1-\alpha/2}]$$

$$Z \sim \mathcal{N}(0,1)$$

Essence of parametric criterions

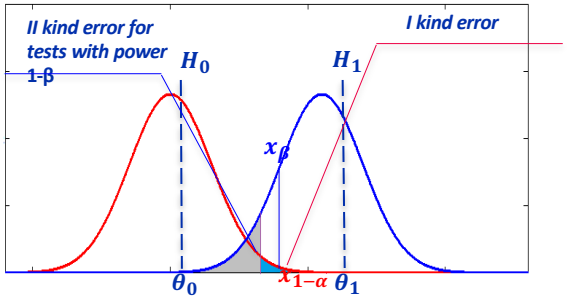


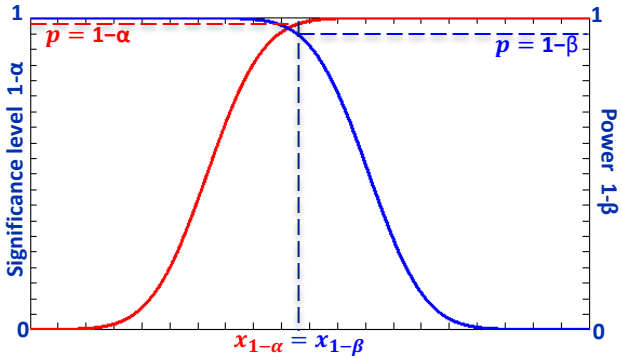
We investigate Θ sample of parameters, that describe distribution of statistical population

$$F(x) = F(x, \Theta)$$

Is it true that $\Theta^* = \Theta_0$ (null hypothesis);
or $\Theta^* = \Theta_1$ (alternative hypothesis)?

Errors of I and II kind





Errors for interval estimation (probability 1-β): *miss* true value from β%- confidence interval

Errors for hypothesis testing:

А) *I kind* (probability α): reject true hypothesis

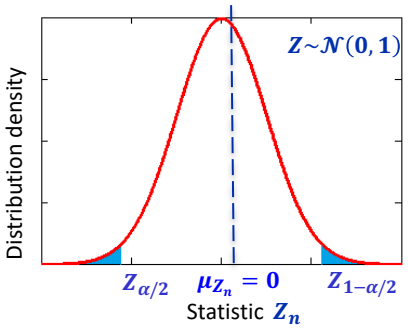
Б) *II kind* (probability γ): accept wrong hypothesis

Building the most powerful tests :

$$1 - \gamma \rightarrow max$$

for fixed significance α

Test for the mean value

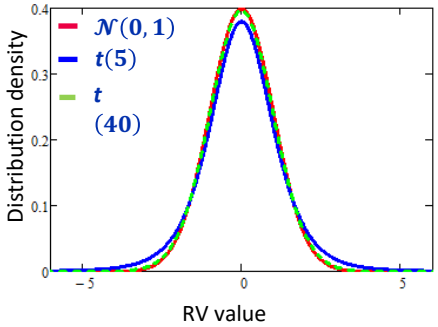


Null hypothesis: $\bar{x} = \mu$

For large samples:

Statistic: $Z_N = \left| \frac{\bar{x} - \mu}{s/\sqrt{N}} \right|$

Test: $Z_N > |Z_{1-\alpha}| \quad Z \sim \mathcal{N}(0,1)$



For small samples (N<30):

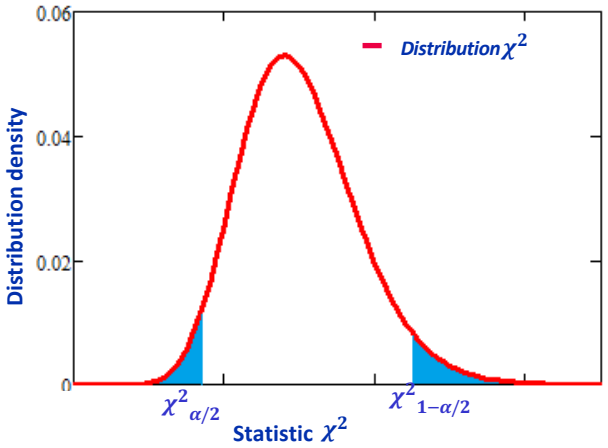
Statistic: $t_N = \sqrt{N-1}\,\frac{|\bar{x} - \mu|}{s}$

Test: $t_N > |t|_{1-\alpha,N-1}$

Test for variance

УНИВЕРСИТЕТ ИТМО

Null hypothesis: $s^2 = \sigma^2$



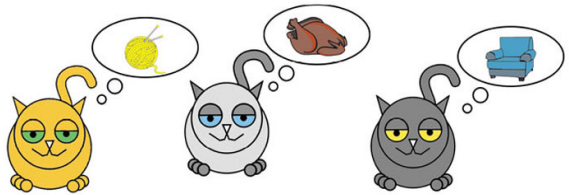
Statistic:

$$\chi_N{}^2 = \frac{(N-1)s^2}{\sigma^2}$$

Test:

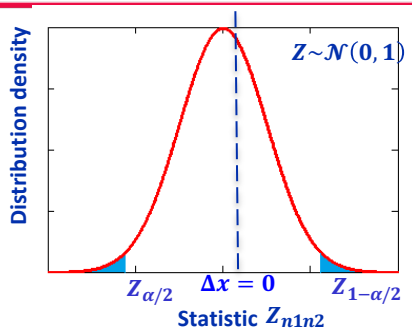
$$\chi_N{}^2 \in \left[0, \chi^2{}_{\frac{\alpha}{2},N-1}\right] \cup \left[\chi^2{}_{1-\frac{\alpha}{2},N-1}, \infty\right]$$

$\chi^2{}_{\frac{\alpha}{2},N-1}$ - quantile of χ^2 distribution



ITMO more than a UNIVERSITY

Test for mean and variance comparison



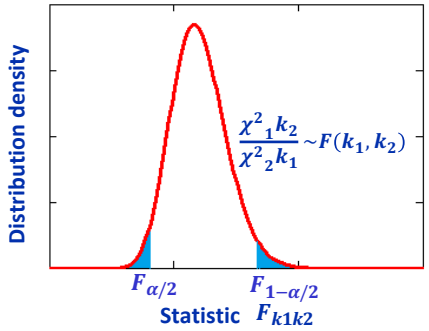
Comparison of mean values:

Statistic:

$$Z_{NM} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{{s_1}^2}{N} + \frac{{s_2}^2}{M}}}$$

Test:

$$Z_{NM} > |Z|_{1-\alpha}$$



Comparison of variances:

Statistic:

$$F_{NM} = \frac{{s_1}^2}{{s_2}^2}$$

Test:

$$F_{NM} \in \left[0, F_{\frac{\alpha}{2},N-1,M-1}\right] \cup \left[F_{1-\frac{\alpha}{2},N-1,M-1}, \infty\right]$$

F - **Fisher distribution**

$$F_{NM} \in \left[0, F_{\frac{\alpha}{2},N-1,M-1}\right] \cup \left[F_{1-\frac{\alpha}{2},N-1,M-1}, \infty\right]$$

Thanks!

www.ifmo.ru

ITMOre than a UNIVERSITY