УНИВЕРСИТЕТ ИТМО

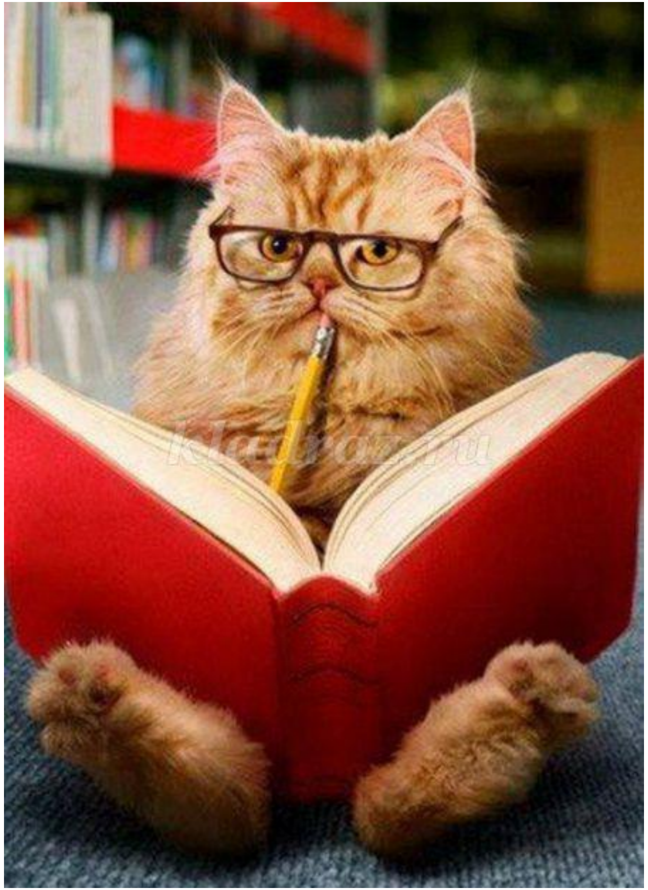
ЗАНЯТИЕ 3

ТЕОРИЯ И ПРАКТИКА СТАТИСТИЧЕСКИХ ВЫВОДОВ

Раздел 1. Генеративная статистика и метод Монте-Карло

Александр Валерьевич Бухановский, Анна Владимировна Калюжная

Понятие статистического вывода



Статистический вывод:
использование выборочных данных для получения и формализации знаний о свойствах генеральной совокупности

- Оценить количественные характеристики генеральной совокупности (вопрос «насколько точно?»)
- Выбрать и настроить модель описания генеральной совокупности (вопрос «подходит или нет?»)
- Валидировать модель, т.е. оценить ее

Метод Монте-Карло

Генеративный (регенеративный) подход: исследование статистических свойств на основе «размножения» заданной выборки

Механизмы генерации:

- На основе априорных знаний о свойствах генеральной совокупности
- На основе действий с выборкой

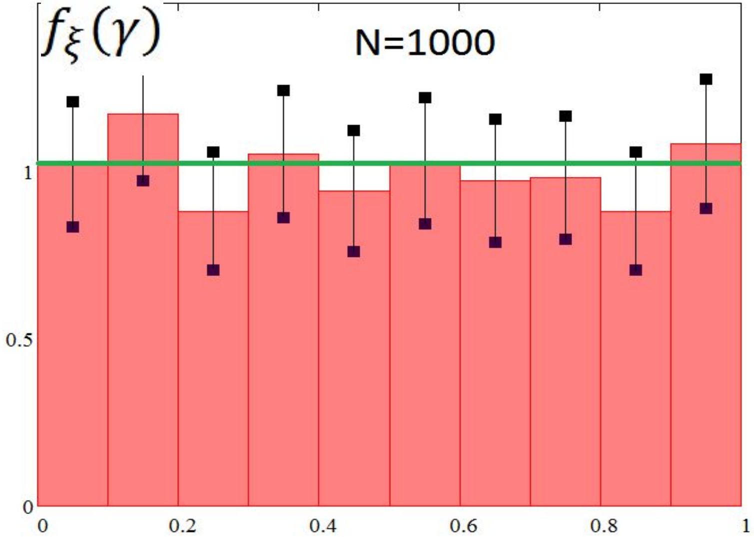
Метод Монте-Карло (метод статистических испытаний): синтез процессов и явлений с использованием случайных чисел

Случайные числа

Значения γ равномерно распределенной случайной величины

в инте

$$f_\xi(\gamma) = \begin{cases} 1, \gamma \in [0,1] \\ 0, \gamma \notin [0,1] \end{cases}$$



Псевдослучайные числа (ПСЧ): на основе детерминированного алгоритма (заданный период повторяемости)

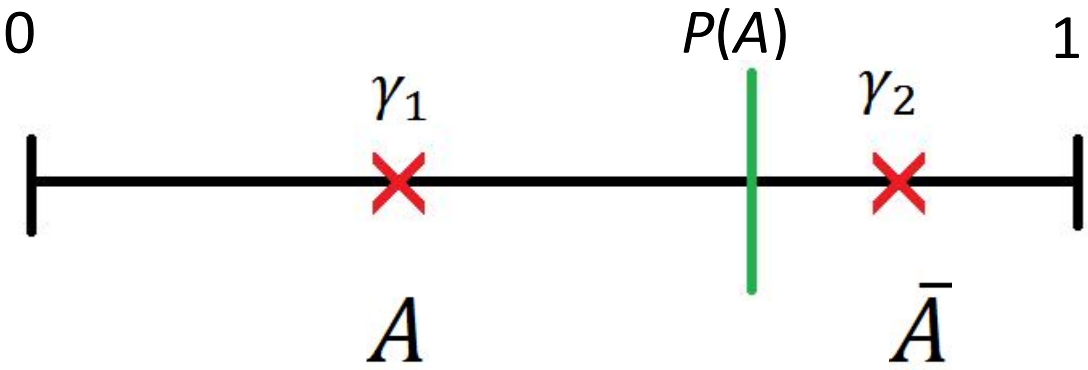
Случайные числа (СЧ): на основе физических механизмов или

Моделирование случайных событий

Геометрическая интерпретация случайного события заданной вероятности $P(A)$ - ***имитация попадания*** точки в интервал $[0, P(A)]$

Выбор варианта реализации события по значению γ:

$$\begin{cases} A, \gamma \le P(A) \\ \bar{A}, \gamma > P(A) \end{cases}$$



Моделирование дискретных величин

Закон распределения (вероятности набора событий)

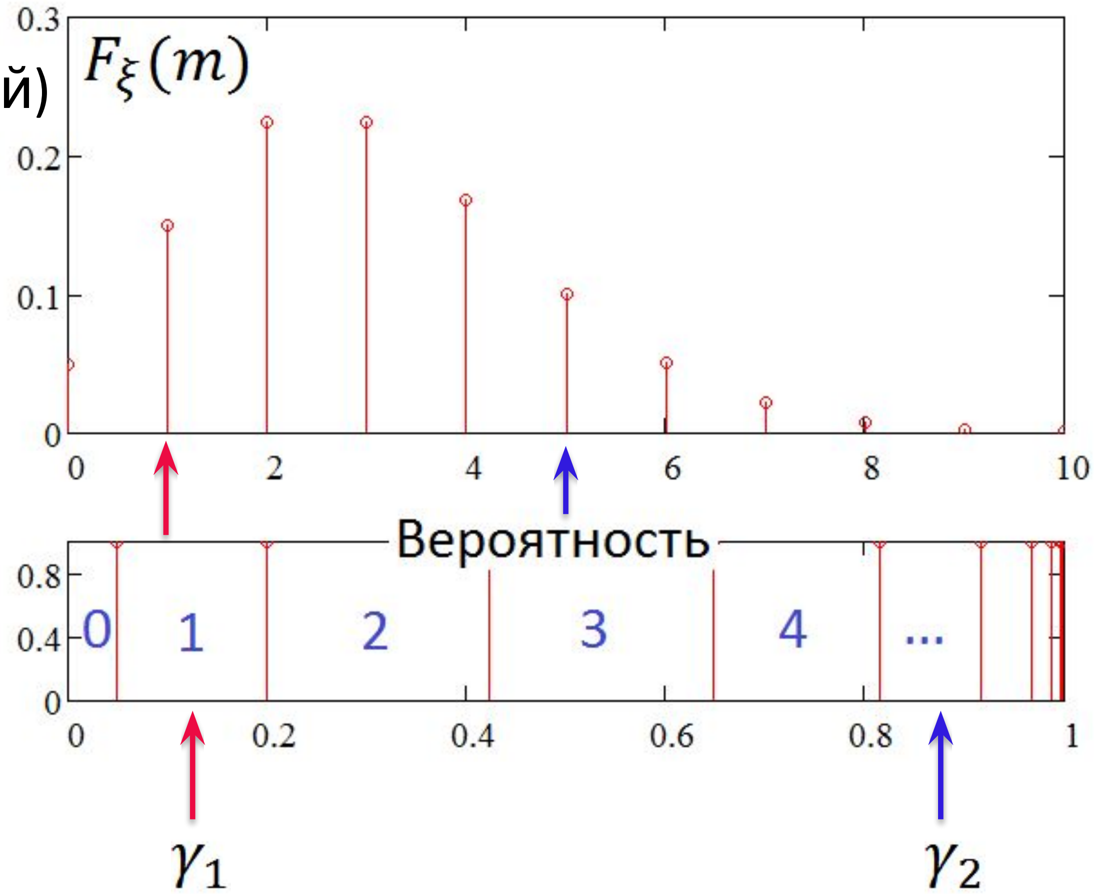
$$F_\xi(m) = P(\xi = x_m), \quad m \in \mathbb{N}$$

Выбор реализации события

$$A_m\colon \xi = x_m$$

при попадании точки в интервал

$$\gamma \in \left[\sum_{k=0}^{m} F_\xi(k), \sum_{k=0}^{m+1} F_\xi(k)\right]$$



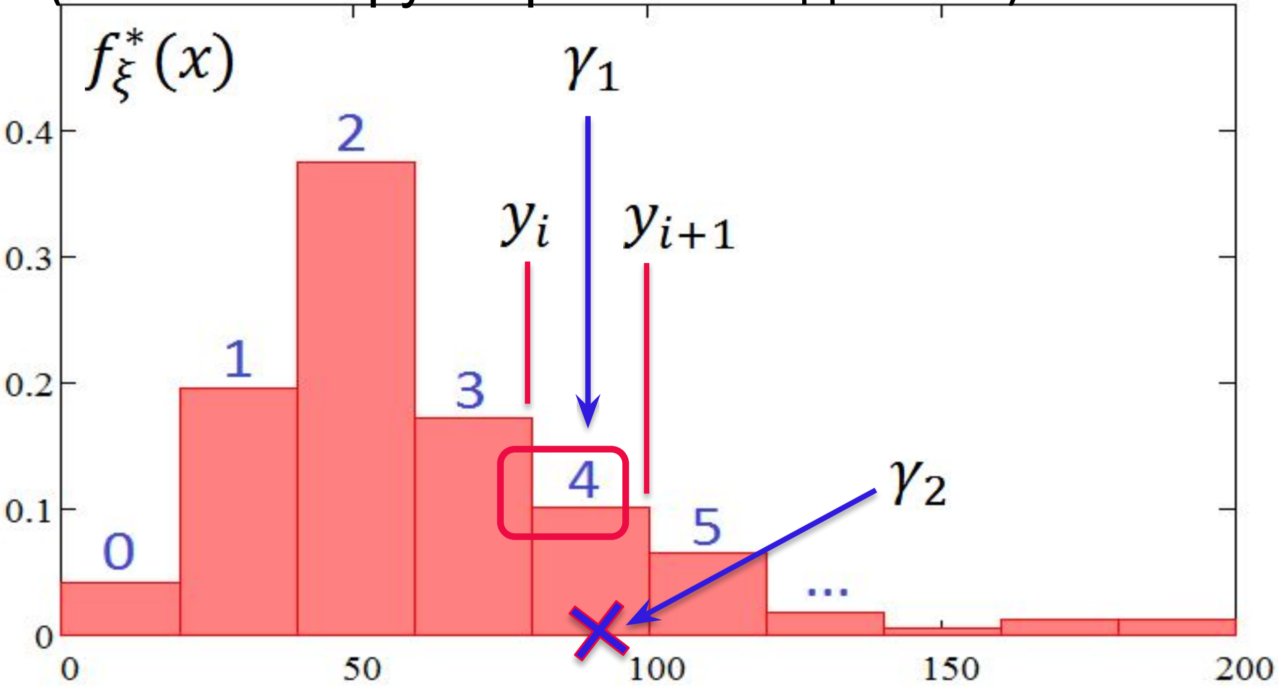
Моделирование непрерывных величин (1/3)

УНИВЕРСИТЕТ ИТМО

Дискретно-равномерное приближение
(на основе группированных данных)

1. Выбор номера столбца i
2. Выбор значения x внутри столбца

$$x = y_i + \gamma(y_{i+1} - y_i)$$



Моделирование непрерывных величин (2/3)

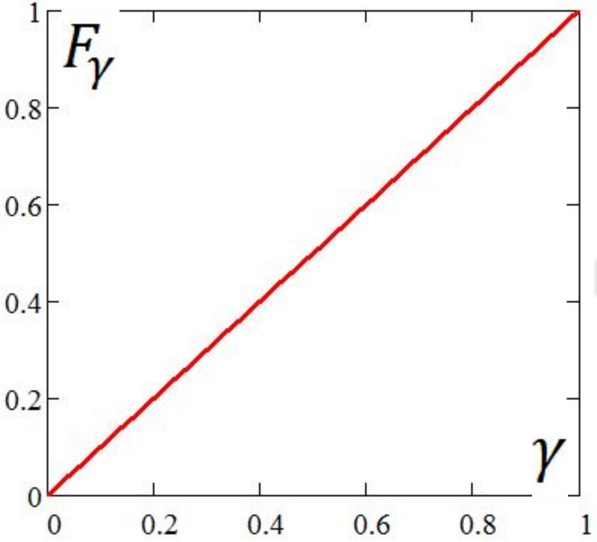
Метод обратной функции

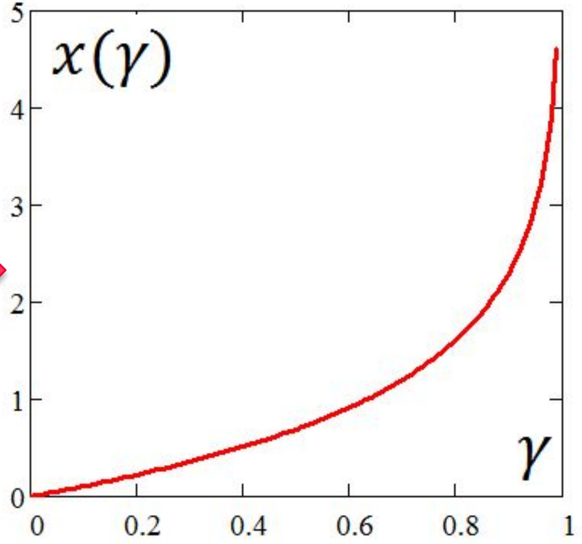
(на основе преобразования квантилей)

$$\boxed{x = F_\xi^{-1}(\gamma)}$$

Для плотности распределения (напр., ядерной оценки)

$$x: \quad \int_{-\infty}^{x} f_\xi(y)dy = \gamma$$





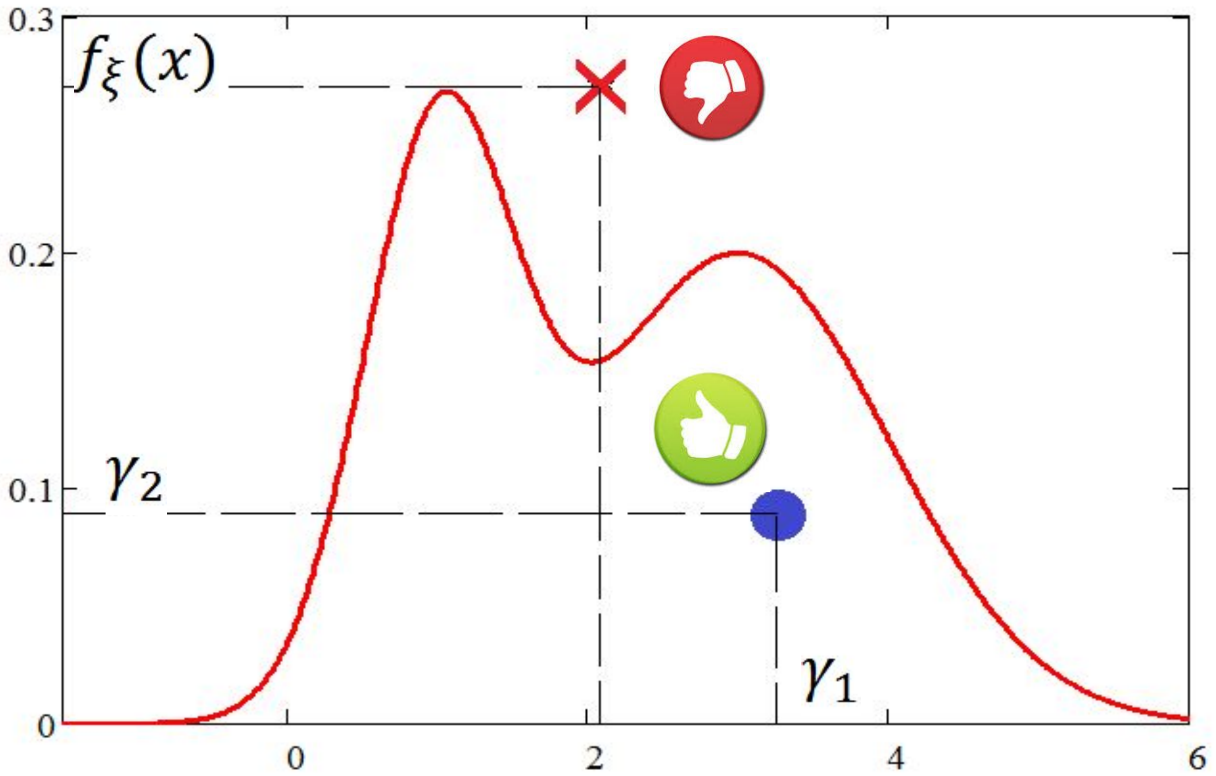
Моделирование непрерывных величин (3/3)

$$x = \gamma_1$$

при условии, что $f_\xi(\gamma_1) \geq \gamma_2$

Геометрический метод

(на основе плотности распределения)



Моделирование гауссовых случайных величин

Нормированное распределение Гаусса **N(0,1)**

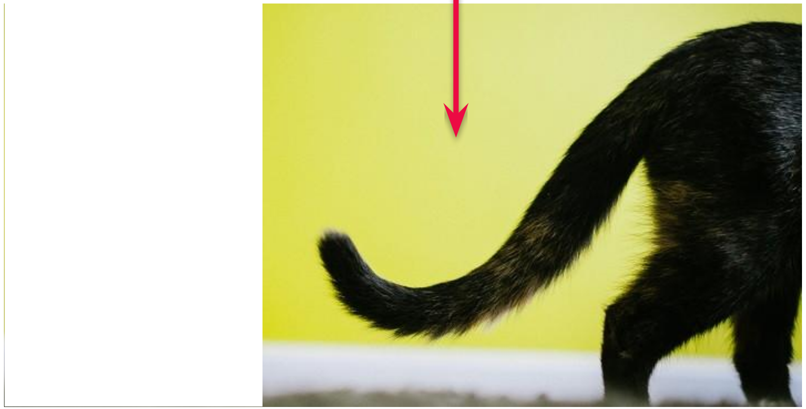
1) На основе **центральной предельной теоремы**:

$$x = \sum_{k=1}^{\boxed{12}} \gamma_i - 6$$

2) На основе **преобразования Бокса-Мюллера:**

$$x_1 = \sin(2\pi\gamma_1)\sqrt{-2\ln\gamma_2}$$

$$x_2 = \cos(2\pi\gamma_1)\sqrt{-2\ln\gamma_2}$$



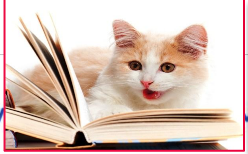
Интервальное оценивание на основе Монте-Карло

Для избранной статистики $\Xi^* = T_P(x_1, \dots, x_N)$:

- Задать аналитическую модель распределе $F_\xi(x, \Theta)$ или ее оценку $F_\xi^*(x) = F_\xi(x, \Theta^*)$
- Выбрать метод моделирования случайной величины с этим распределением
- Сгенерировать M выборо $\left(x_1^{(j)}, \dots, x_N^{(j)}\right), j = 1 \dots M$
- Рассчитать по каждой выборке точечную оцен Ξ_j^* статист…и Ξ^*
- По выборке $(\Xi_1^*, \dots, \Xi_M^*)$ определить характерные параметры, в том числе границы β%-доверительного интервала:

$$I_1^{(\beta)} = \Xi^*_{(\lfloor M(1-\beta)/2 \rfloor + 1)} \qquad I_2^{(\beta)} = \Xi^*_{(\lfloor M(1+\beta)/2 \rfloor + 1)}$$

Задача об отборе на выставку кошек (1/2)



Найти интервальные оценки математического ожидания и дисперсии для: (а) выборки из всех 109 кошек, (б) для случайно отобранных 10

Чтение исходных данных - чистая выборка (кошки)

$X := READPRN("109\_Cats.dat")$ $N := rows(X)$ $i := 0..N-1$ $N = 109$

$u\beta := 1.96$ Квантиль распределения Гаусса для 95% интервала $\beta := 0.95$

Точечные оценки по выборке

$Mx := mean(X)$ $Mx = 5.806$ Среднее значение

$Sx := stdev(X)$ $Sx = 0.425$ Выборочное СКО

$Dx := var(X)$ $Dx = 0.181$ Выборочная дисперсия

Доверительные интервалы для выборочных средних и дисперсий - расчет по асимптотическим формулам

95%-доверительные интервалы для выборочного среднего

$$IMniz := Mx - u\beta\cdot\frac{Sx}{\sqrt{N}} \qquad IMver := Mx + u\beta\cdot\frac{Sx}{\sqrt{N}}$$

$IMniz = 5.726$ $IMver = 5.886$

95%-доверительные интервалы для выборочной дисперсии

$$IDniz := \frac{Dx\cdot(N-1)}{qchisq(0.975, N)} \qquad IDver := \frac{Dx\cdot(N-1)}{qchisq(0.025, N)}$$

$IDniz = 0.14$ $IDver = 0.238$

Доверительные интервалы для выборочных средних и дисперсий - расчет методом Монте-Карло

$M := 10000$ Число испытаний метода Монте-Карло $j := 0..M-1$

$k := 0..11$

$$Y_{i,j} := Mx + Sx\cdot\left(\sum_k rnd(1) - 6\right)$$

Генератор M выборок гауссовых случайных чисел со средним Mx и СКО Sx, каждая из N значений

Расчет средних и дисперсий для каждой из выборок

$Mx\_MC_j := mean\left(Y^{\langle j\rangle}\right)$ $Dx\_MC_j := var\left(Y^{\langle j\rangle}\right)$

Расчет границ доверительных (толерантных) интервалов по выборкам Mx_MC и Dx_MC - как соответствующих квантилей

$no(p) := floor(M\cdot p)$ $Mx\_MC\_sort := sort(Mx\_MC)$ $Dx\_MC\_sort := sort(Dx\_MC)$

95%-доверительные интервалы для выборочного среднего

$Mx\_MC\_sort_{no(0.025)} = 5.727$ $Mx\_MC\_sort_{no(0.97)} = 5.882$

95%-доверительные интервалы для выборочной дисперсии

$Dx\_MC\_sort_{no(0.025)} = 0.136$ $Dx\_MC\_sort_{no(0.975)} = 0.23$

Задача об отборе на выставку кошек (2/2)

Чтение исходных данных - чистая выборка (кошки)

$X := \mathrm{READPRN}("10\_SmallSampleCats.dat")$ $N := \mathrm{rows}(X)$ $i := 0..N-1$ $N = 10$

$u\beta := 1.96$ Квантиль распределения Гаусса для 95% интервала

Точечные оценки по выборке

$Mx := \mathrm{mean}(X)$ $Mx = 5.968$ Среднее значение

$Sx := \mathrm{stdev}(X)$ $Sx = 0.39$ Выборочное СКО

$Dx := \mathrm{var}(X)$ $Dx = 0.152$ Выборочная дисперсия

Доверительные интервалы для выборочных средних и дисперсий - расчет по асимптотическим формулам

95%-доверительные интервалы для выборочного среднего

$IMniz := Mx - u\beta \cdot \frac{Sx}{\sqrt{N}}$ $IMver := Mx + u\beta \cdot \frac{Sx}{\sqrt{N}}$ $IMniz = 5.726$ $IMver = 6.21$

95%-доверительные интервалы для выборочной дисперсии

$IDniz := \frac{Dx \cdot (N-1)}{\mathrm{qchisq}(0.975, N)}$ $IDver := \frac{Dx \cdot (N-1)}{\mathrm{qchisq}(0.025, N)}$ $IDniz = 0.067$ $IDver = 0.421$

Доверительные интервалы для выборочных средних и дисперсий - расчет методом Монте-Карло

$M := 10000$ Число испытаний метода Монте-Карло $j := 0..M-1$

$k := 0..11$

$Y_{i,j} := Mx + Sx \cdot \left(\sum_k \mathrm{rnd}(1) - 6\right)$ Генератор M выборок гауссовых случайных чисел со средним Mx и СКО Sx, каждая из N значений

Расчет средних и дисперсий для каждой из выборок

$Mx\_MC_j := \mathrm{mean}\left(Y^{\langle j \rangle}\right)$ $Dx\_MC_j := \mathrm{var}\left(Y^{\langle j \rangle}\right)$

Расчет границ доверительных (толерантных) интервалов по выборкам Mx_MC и Dx_MC - как соответствующих квантилей

$no(p) := \mathrm{floor}(M \cdot p)$ $Mx\_MC\_sort := \mathrm{sort}(Mx\_MC)$ $Dx\_MC\_sort := \mathrm{sort}(Dx\_MC)$

95%-доверительные интервалы для выборочного среднего

$Mx\_MC\_sort_{no(0.025)} = 5.73$ $Mx\_MC\_sort_{no(0.975)} = 6.213$

95%-доверительные интервалы для выборочной дисперсии

$Dx\_MC\_sort_{no(0.025)} = 0.04$ $Dx\_MC\_sort_{no(0.975)} = 0.287$

Для малых выборок асимптотические формулы некорректны, а метод Монте-Карло – без

Непараметрические методы: бутстреп

Формирование *M* *псевдовыборок*

$$\left(x_1^{(j)}, \dots, x_N^{(j)}\right), j = 1 \dots M$$

на основе случайного *выбора с возвращением* из исходной выборки:

А Б В Г Д Е

Непараметрические методы: джекнайф

Формирование *N* **подвыборок**

$$\left(x_1^{(j)}, \dots, x_{N-1}^{(j)}\right), j = 1 \dots N$$

путем ***выкалывания*** одного из членов исходной выборки:

А Б В Г Д Е

Методическая проблема джекнайфа и бутстрепа: ***смещенность*** оценок

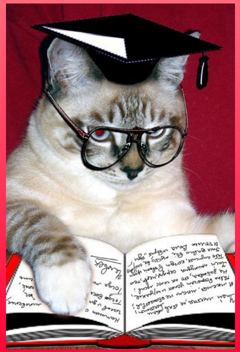
ITsMOre than a UNIVERSITY

Обобщение: генеративная статистика

УНИВЕРСИТЕТ ИТМО

- **Метод Монте-Карло:** универсальный инструмент статистического оценивания, когда аналитика не справляется
- **Методы моделирования случайных величин:** могут использовать как параметрические, так и непараметрические оценки распределений
- **Интервальное оценивание** на основе метода Монте-Карло: просто, надежно, но чувствительно к модели распределения
- **Методы бутстреп и джекнайф:** не зависят от распределения, но оценки могут быть смещены

Автор признателен всем котикам (и хозяевам) за возможность некоммерческого использования их изображений, размещенных в публичном Интернете



Александр Валерьевич Бухановский, Анна Владимировна Калюжная
boukhanovsky@mail.ifmo.ru

ITsMOre than a UNIVERSITY

УНИВЕРСИТЕТ ИТМО

ЗАНЯТИЕ 3

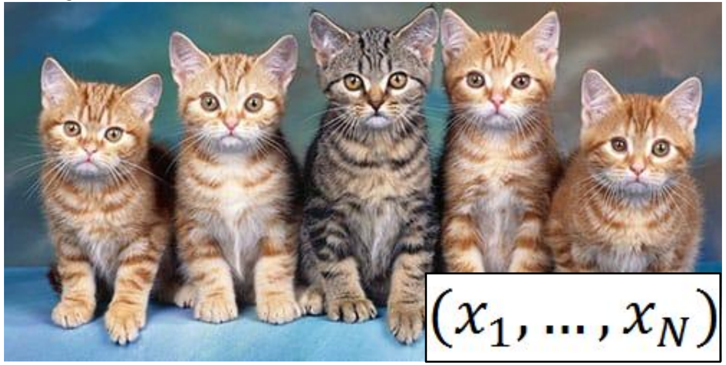
ТЕОРИЯ И ПРАКТИКА СТАТИСТИЧЕСКИХ ВЫВОДОВ

Раздел 2. Проверка гипотез. Непараметрические статистические критерии

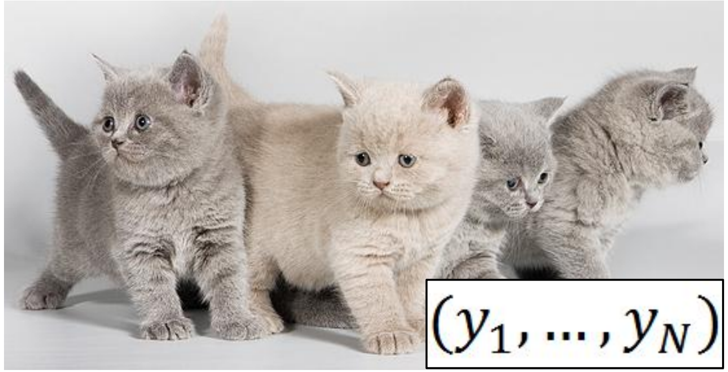
Александр Валерьевич Бухановский, Анна Владимировна Калюжная

Проверка статистических гипотез

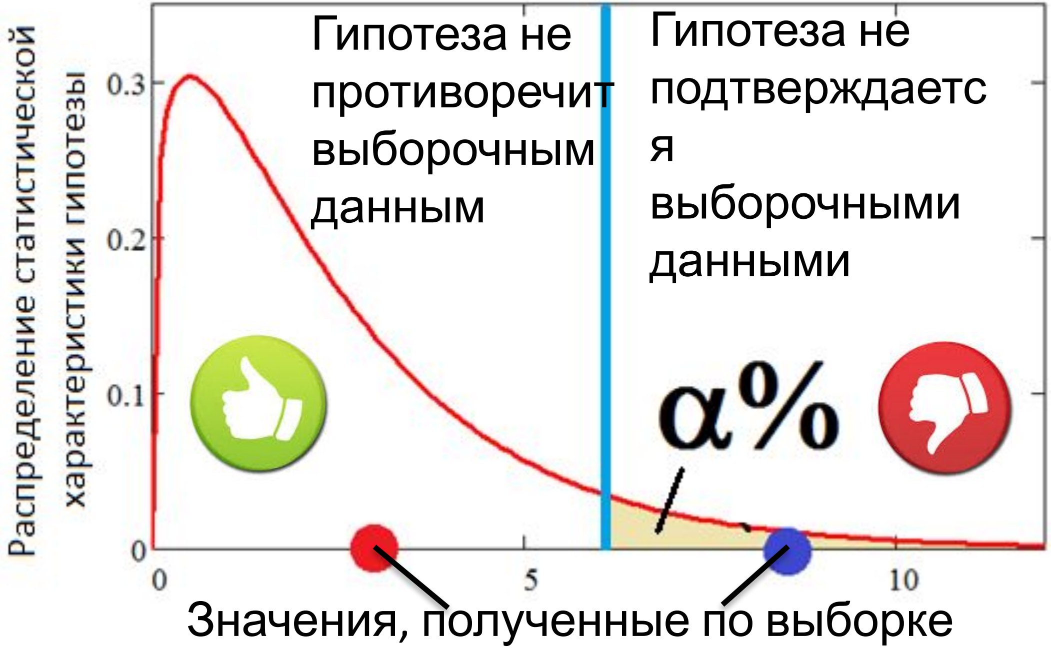
Нулевая (базовая) гипотеза



? II ?

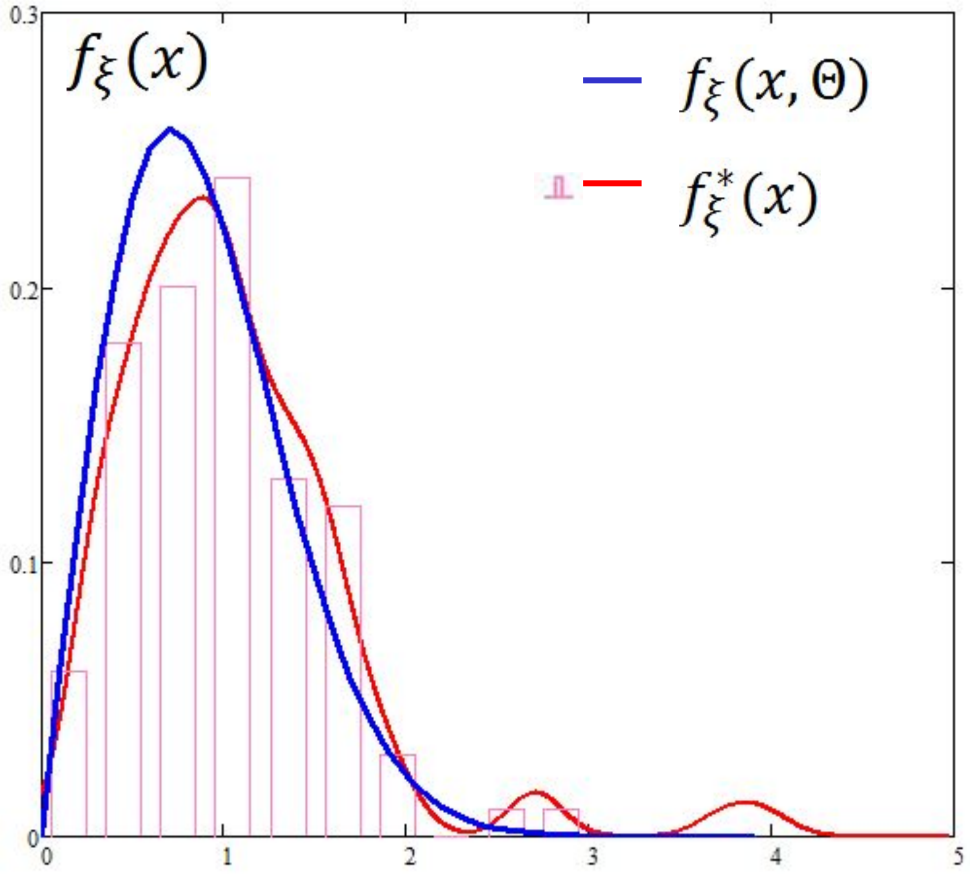


Статистический критерий: механизм проверки нулевой гипотезы путем сравнения выборочных данных с теоретическим «эталоном»



Гипотеза о виде распределения (пример)

УНИВЕРСИТЕТ ИТМО



Критерий согласия

Базовая гипотеза H_0 :

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

Альтернатива H_1 :

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

Теоретическая характеристика

$$T = \mathfrak{G}[f_\xi]$$

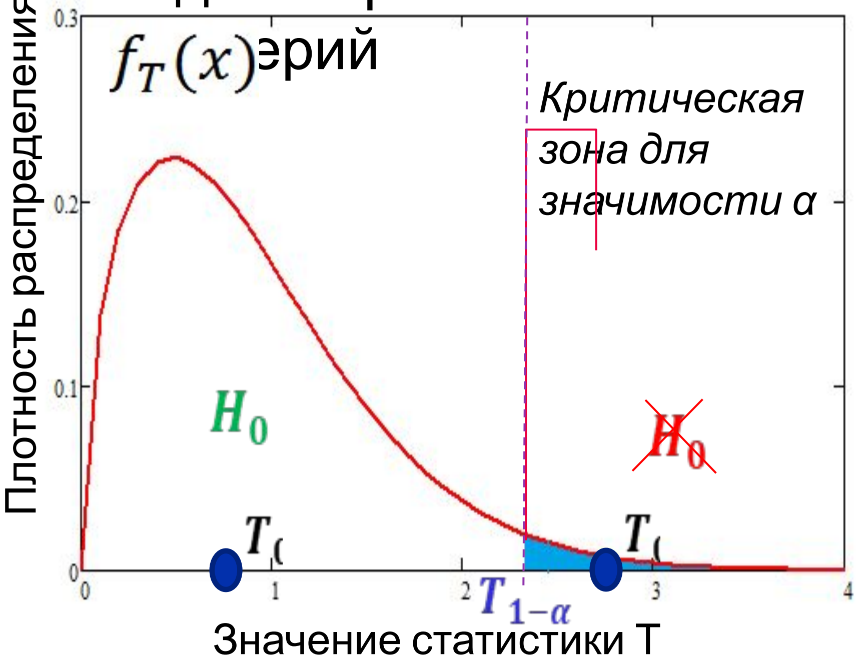
Выборочная

ра

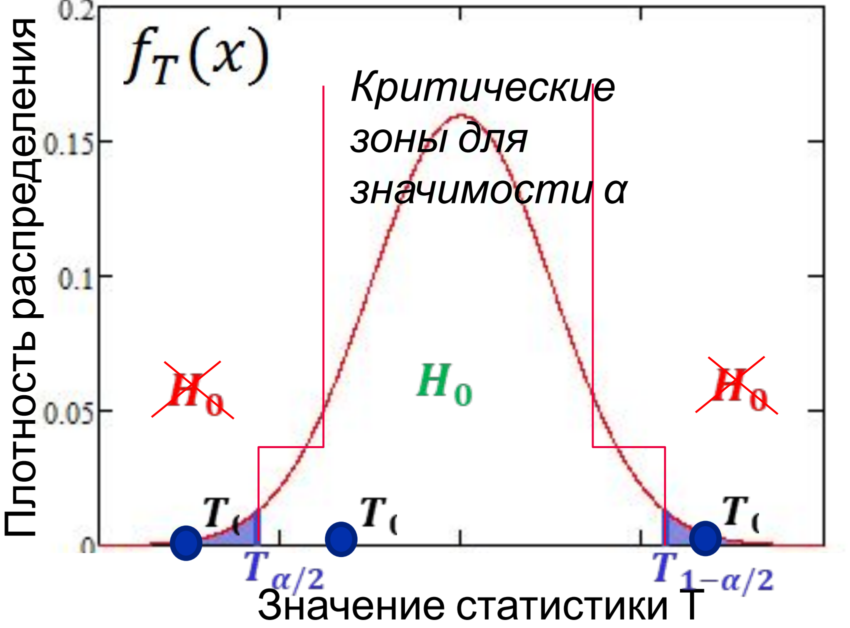
$$T_0 = \mathfrak{F}[f_\xi, f_\xi^*(x_i)]$$

Критерии: сравнение характеристик

Односторонний критерий

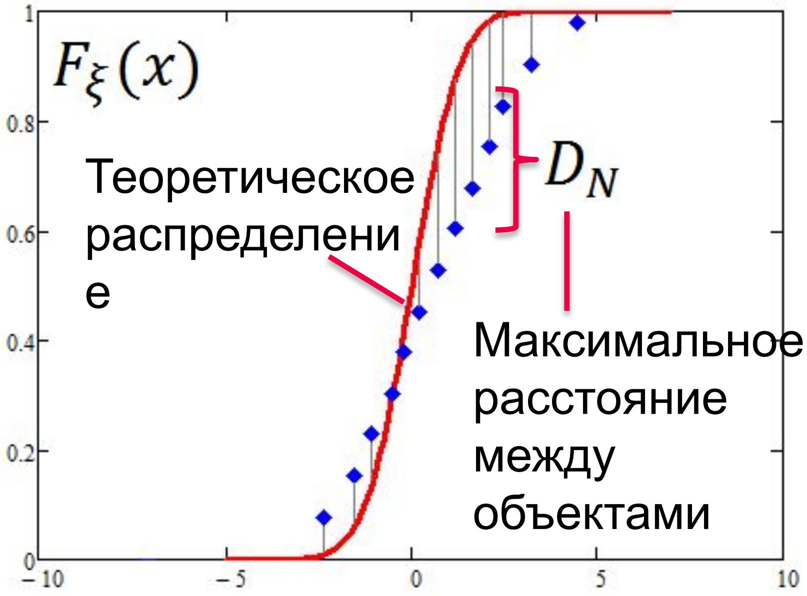


Двусторонний критерий



Ошибка проверки непараметрических гипотез
(*ошибка I рода*) – с вероятностью α отвергнуть верную

Критерий Колмогорова



Проверка гипотезы **о виде (модели)** распределения

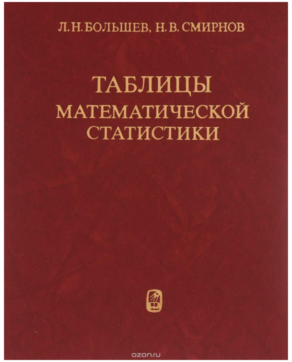
Статистическая характеристика:

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

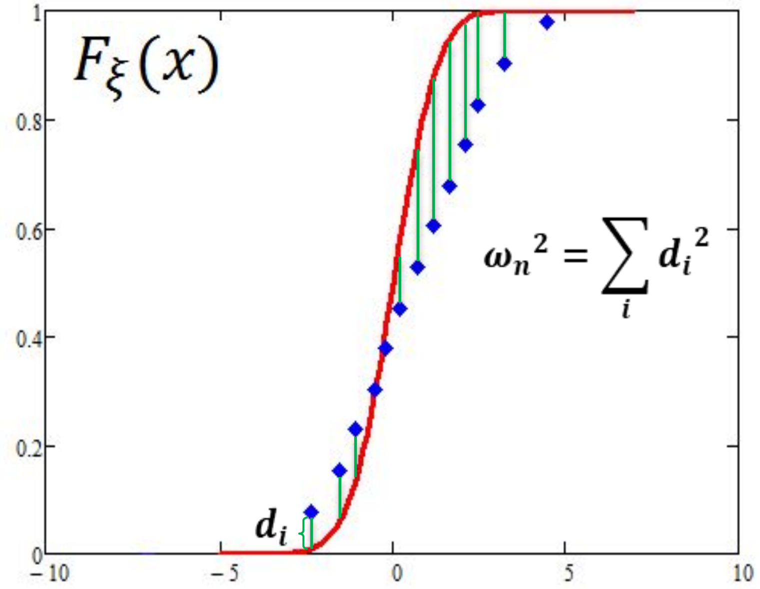
Критерий проверки (***отрицани*** H_0):

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

$T_{1-\alpha}$ - квантиль распределения Колмогорова



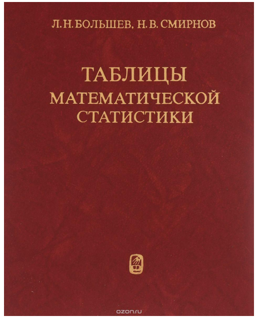
Критерий ω^2 (Крамера-вон Мизеса-Смирнова)



Проверка гипотезы **о виде (модели)** распределения

Статистическая характеристика:

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

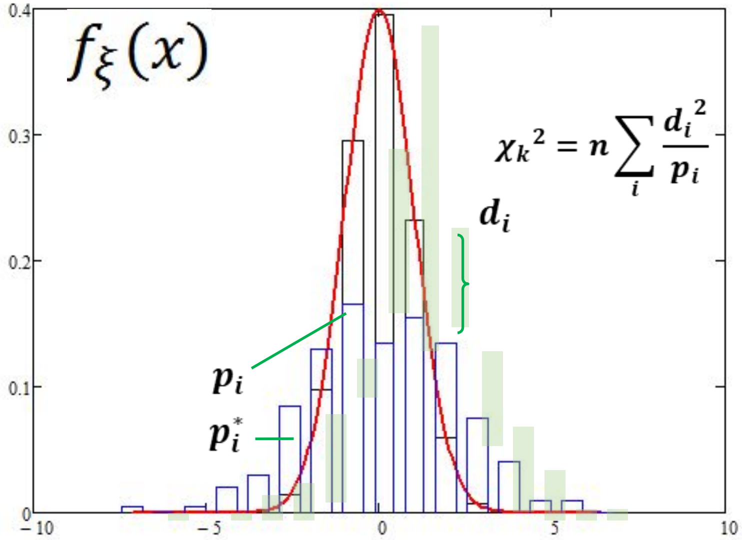


Критерий проверки (***отрицан*** H_0):

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

$T_{1-\alpha}$ - квантиль табулированного

Критерий χ^2 (Пирсона)



Проверка гипотезы **о виде (модели)** распределения *по группированным данным*

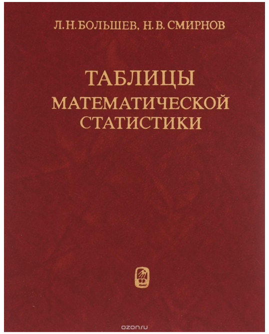
Статистическая характеристика :

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

Критерий проверки (***отрицани*** H_0):

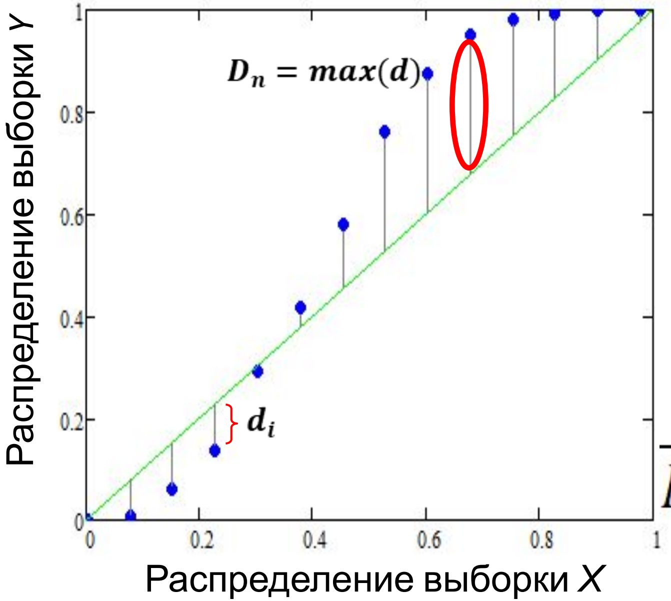
$$\chi_N^2 \geq T_{1-\alpha}$$

$T_{1-\alpha}$ - квантиль распределения χ_{N-1}^2



Критерии для проверки однородности

Вероятностный



Базовая гипотеза $H_0 : f_\xi^*(x) = g_\xi^*(x)$

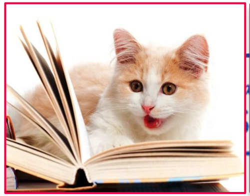
Критерий Колмогорова

См

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

Критерий ω^2 Розенблатта

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



1.4

Правила проверки
–
как для критериев

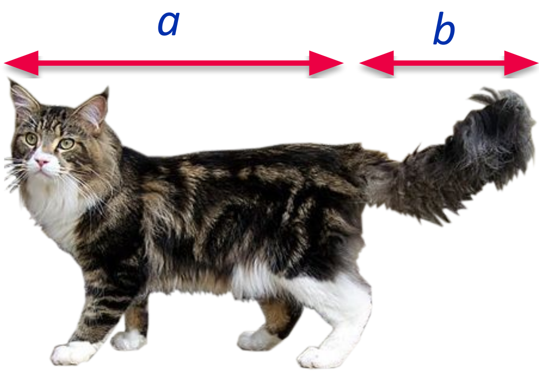
Задача о поддельных мейн-кунах (1/3)

УНИВЕРСИТЕТ ИТМО

В одном питомнике разводят крупных котов и продают как мейн-кунов. Однако ассоциация заводчиков мейн-кунов засомневалась в том, что эти коты действительно принадлежат породе, а не являются смесью с норвежской лесной кошкой. Для того, чтобы принять или отклонить гипотезу о принадлежности котов из питомника породе мейн-кун, заводчики решили сравнить распределения характерного параметра кошек между собой. В качестве параметров для сравнения могут быть выбраны отношения длин тел животных, хвостов или хвоста к длине тела.

Известно, что в средние длина тел котов норвежской породы и мейн-кунов совпадает и составляет 47 - 50 см. Известно также, что отношение длины хвоста к телу у породистых норвежских кошек должно быть больше 0,9, а у породистых мейн-кунов больше 3/5. Для чистоты эксперимента сравнение решили проводить только на взрослых котах, которых в питомнике оказалось 30. Из базы ассоциации мейн-кунов для сопоставления были выбраны 43 самых породистых кота.

Однако, стоит учесть, что если кошки из питомника являются смесью пород, тогда ими могут быть унаследованы как признаки



$x = a/b$

Задача о поддельных мейн-кунах (2/3)

1) Выборки длин тел мейн-кунов и котов из питомника:

$BLM := M^{\langle 0 \rangle}$ $BLF := F^{\langle 0 \rangle}$

2) Выборки длин хвостов мейн-кунов и котов из питомника:

$TLM := M^{\langle 1 \rangle}$ $TLF := F^{\langle 1 \rangle}$

3) Выборки отношений длин хвостов к телам мейн-кунов и котов из питомника:

$RM := M^{\langle 2 \rangle}$ $RF := F^{\langle 2 \rangle}$

M := READPRN("Meinkun_data.dat") F := READPRN("Farm_cats_data.dat")

M - мейн-куны **F** - коты из питомника

$m := rows(M)$ $n := rows(F)$

$m = 43$ $n = 30$

$j := 0..m-1$ $i := 0..n-1$

Получаем эмпирические функции распределения на основе ядерных оценок с гауссовым ядром:

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

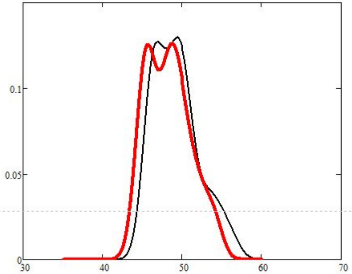
Для длин тел:

$hBLM := 1.1$ $hBLF := 1.1$

$$fBLM(y) := \frac{1}{m\cdot hBLM}\cdot\sum_j K\_gauss\left(\frac{y - BLM_j}{hBLM}\right) \qquad fBLF(y) := \frac{1}{n\cdot hBLF}\cdot\sum_i K\_gauss\left(\frac{y - BLF_i}{hBLF}\right)$$

$d := 0.1$

$y := 35, 35 + d..60$



$z := 35, 35 + d..60$

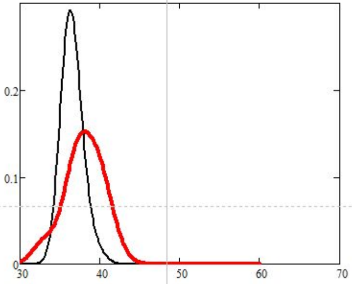
Для длин хвостов:

$hTLM := 1.1$ $hTLF := 1.3$

$$fTLM(y) := \frac{1}{m\cdot hTLM}\cdot\sum_j K\_gauss\left(\frac{y - TLM_j}{hTLM}\right) \qquad fTLF(y) := \frac{1}{n\cdot hTLF}\cdot\sum_i K\_gauss\left(\frac{y - TLF_i}{hTLF}\right)$$

$d := 0.1$

$yt := 30, 30 + d..60$



$zt := 30, 30 + d..60$

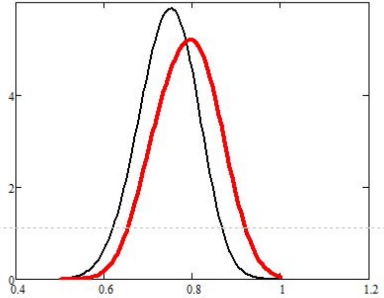
Для отношения длин тел к хвостам:

$hRM := 0.05$ $hRF := 0.05$

$$fRM(y) := \frac{1}{m\cdot hRM}\cdot\sum_j K\_gauss\left(\frac{y - RM_j}{hRM}\right) \qquad fRF(y) := \frac{1}{n\cdot hRF}\cdot\sum_i K\_gauss\left(\frac{y - RF_i}{hRF}\right)$$

$dr := 0.005$

$yr := 0.5, 0.5 + dr..1$



$zr := 0.5, 0.5 + dr..1$

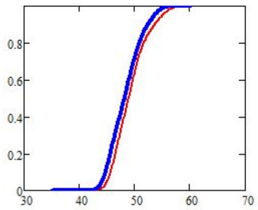
Задача о поддельных мейн-кунах (3/3)

Проверка на однородность выборок с помощью критерия Смирнова:

Квантиль распределения Колмогорова для уровня значимости 5%:

$Dcr := 1.36$

$FBLM(z) := \int_{35}^{z} fBLM(y)\,dy \qquad FBLF(z) := \int_{35}^{z} fBLF(y)\,dy$



Для длин тел:

$$Lx := floor\left[\frac{(\max(BLM, BLF) - \min(BLM, BLF))}{d}\right]$$

$Lx = 119$

$p := 0..Lx$

$X(p) := \min(BLM, BLF) + 0.1 \cdot p$

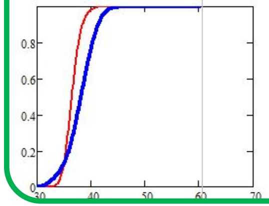
$D_p := |FBLM(X(p)) - FBLF(X(p))|$

$$DD := \sqrt{\frac{(n \cdot m)}{n+m}} \cdot \max(D)$$

$DD = 0.5$

DD < Dcr, поэтому H0 принята

$FTLM(zt) := \int_{30}^{zt} fTLM(yt)\,dyt \qquad FTLF(zt) := \int_{30}^{zt} fTLF(yt)\,dyt$



Для длин хвостов:

$$Lxt := floor\left[\frac{(\max(TLM, TLF) - \min(TLM, TLF))}{d}\right]$$

$Lxt = 100$

$pt := 0..Lxt$

$Xt(pt) := \min(TLM, TLF) + d \cdot pt$

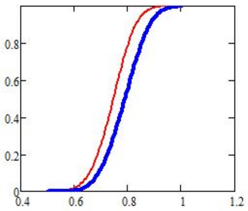
$Dt_{pt} := |FTLM(Xt(pt)) - FTLF(Xt(pt))|$

$$DDt := \sqrt{\frac{(n \cdot m)}{n+m}} \cdot \max(Dt)$$

$DDt = 1.539$

DDt > Dcr, поэтому H0 отвергнута

$FRM(zr) := \int_{0.5}^{zr} fRM(yr)\,dyr \qquad FRF(zr) := \int_{0.5}^{zr} fRF(yr)\,dyr$



Для отношения длин тел к хвостам:

$$Lxr := floor\left[\frac{(\max(RM, RF) - \min(RM, RF))}{dr}\right]$$

$Lxr = 51$

$pr := 0..Lxr$

$Xr(pr) := \min(RM, RF) + dr \cdot pr$

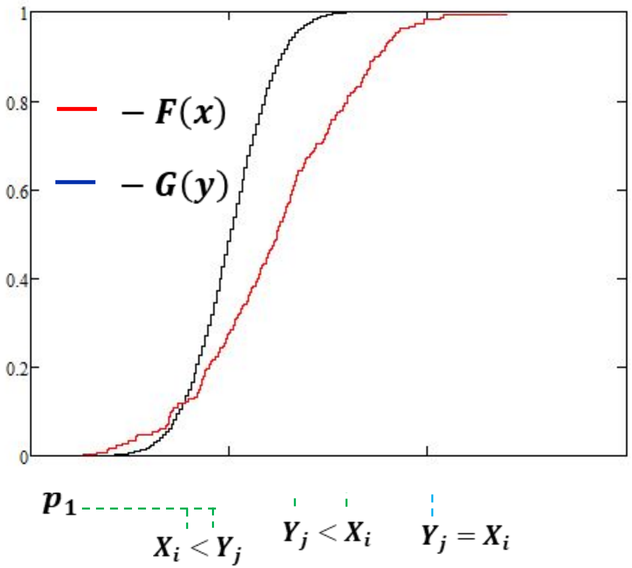
$Dr_{pr} := |FRM(Xr(pr)) - FRF(Xr(pr))|$

$$DDr := \sqrt{\frac{(n \cdot m)}{n+m}} \cdot \max(Dr)$$

$DDr = 1.003$

DDr < Dcr, поэтому H0 принята

Критерий ранговых сумм Вилкоксона



Неоднородность как превалирование

Базовая гипотеза $H_0 \quad F(x) = G(x)$

Альтернативная гипотеза $H_2 \quad F(x) < G(x)$

Выборки $(x_1, \dots, x_N)$ и $(y_1, \dots, y_M)$

Статистическая характеристика:

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

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

Критерий проверки (***отрицание*** H_0):

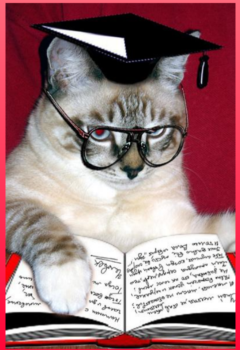
$$U_{st} \notin [Z_{\alpha/2}, Z_{1-\alpha/2}]$$

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

Обобщение: непараметрические критерии

- **Решаемые задачи:** проверка соответствие распределения определенному закону, проверка однородности двух выборок.
- **Все критерии несовершенны:** критерий Колмогорова прост, но чувствителен к выбросам, критерий согласия Пирсона – для группированных данных (все проблемы гистограмм), критерий Крамера-вон Мизеса-Смирнова табулирован не для всех распределений
- **Проверку можно строить по-разному: просто проверять неоднородность,** модифицируя критерии для теоретических распределений (Смирнова, Розенблатта), или оценивать превалирование (критерии Вилкоксона и Манна-Уитни).
- **Проверка гипотез – не панацея:** она может только опровергать,

Автор признателен всем котикам (и хозяевам) за возможность некоммерческого использования их изображений, размещенных в публичном Интернете



Александр Валерьевич Бухановский, Анна Владимировна Калюжная
boukhanovsky@mail.ifmo.ru

ITsMOre than a UNIVERSITY

УНИВЕРСИТЕТ ИТМО

ЗАНЯТИЕ 3

ТЕОРИЯ И ПРАКТИКА СТАТИСТИЧЕСКИХ ВЫВОДОВ

Раздел 3. Проверка гипотез. Параметрические статистические критерии

Александр Валерьевич Бухановский, Анна Владимировна Калюжная

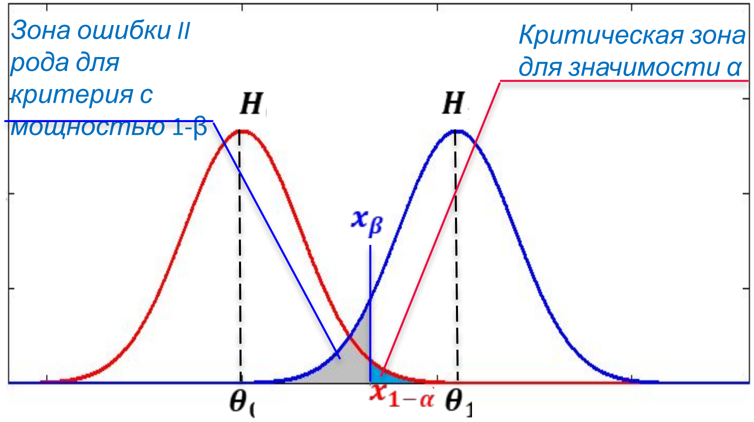
Логика параметрических критериев

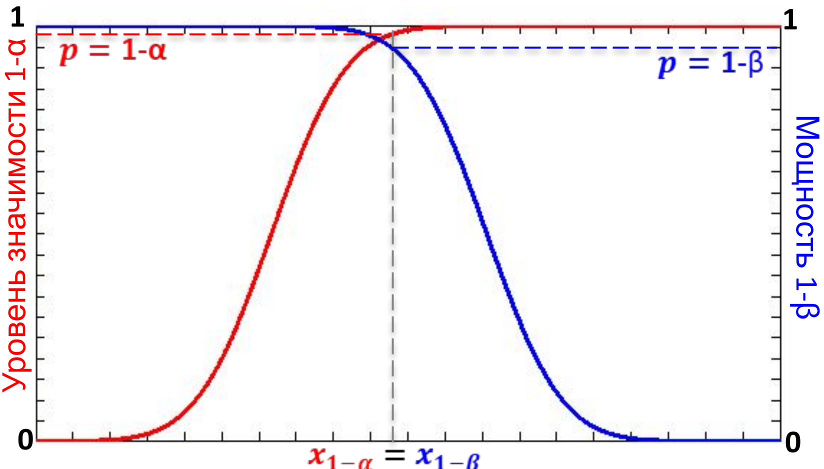
УНИВЕРСИТЕТ ИТМО

Объект исследования – набор параметров Θ, характеризующих модель распределения генеральной совокупности $F(x) = F(x, \Theta)$

А правда ли, что параметр $\Theta^* = \Theta_0$ (базовая гипотеза);
или все-таки парам $\Theta^* = \Theta_1$ (альтернативная гипотеза)?

Ошибки первого и второго рода





Ошибка статистического оценивания (вероятность 1-β): *упустить* истинное значение из β%-доверительного интервала

Ошибки проверки гипотез:

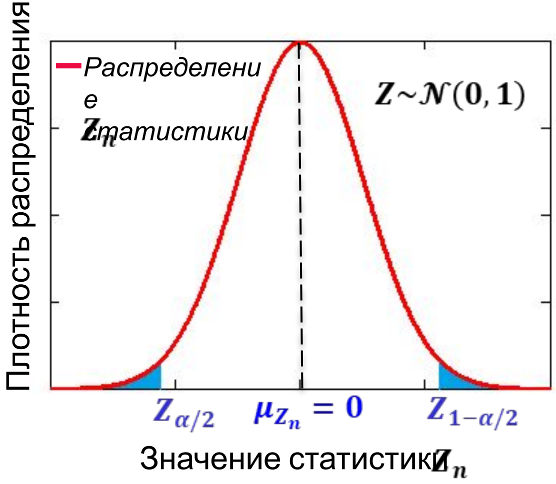
А) *Ошибка I рода* (вероятность α): отвергнуть верную гипотезу

Б) *Ошибка II рода* (вероятность γ): принять неверную гипотезу

Построение наиболее *мощных* (в своем классе) критериев: $\boxed{1 - \gamma \to max}$

при заданном уровне значимости α

Критерий для среднего значения



Базовая гипотеза: $\boxed{\bar{x}=\mu}$

Для больших выборок:

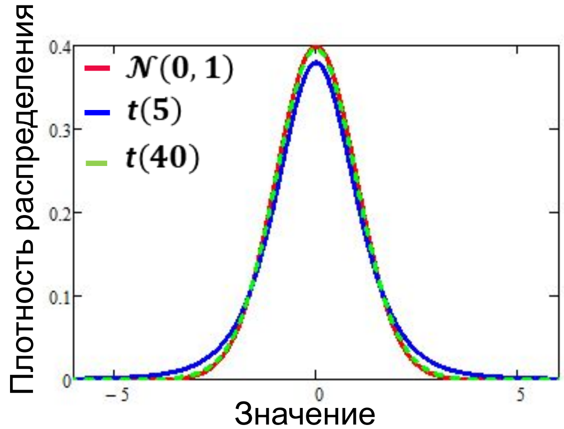
Статистическая характеристика $Z_N=\left|\dfrac{\bar{x}-\mu}{s/\sqrt{N}}\right|$

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

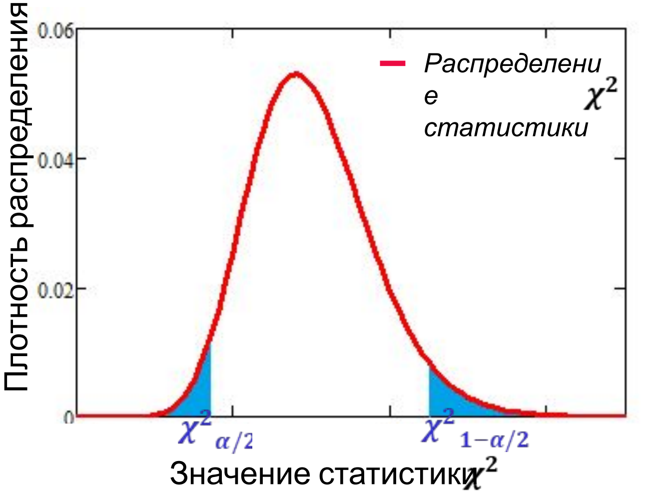
Для малых выборок (N<30):

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

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



χ^2 Критерий для выборочной дисперсии



Базовая гипотеза: $\boxed{s^2 = \sigma^2}$

Статистическая характеристика:

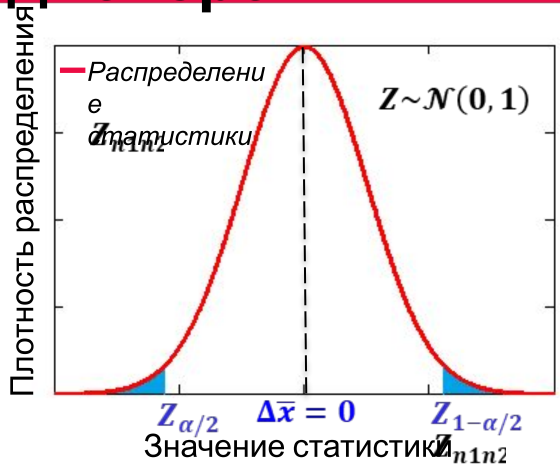
$$\chi_N{}^2 = \frac{(N-1)s^2}{\sigma^2}$$

Критерий проверки (***отрицани*** H_0):

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

$\chi^2{}_{\frac{\alpha}{2},N-1}$ - табличная квантиль распределения χ^2 с *N-1* степенью свободы

Критерий для равенства средних и дисперсий



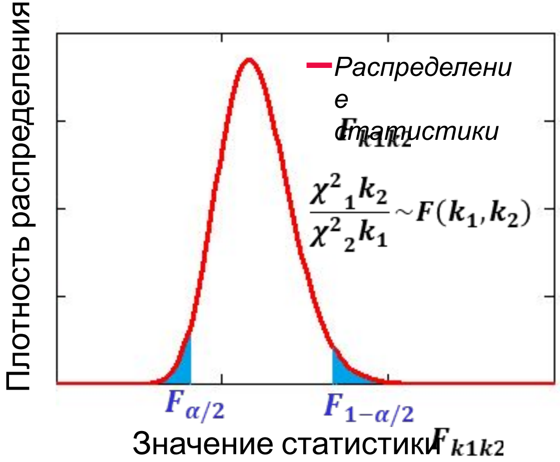
Равенство двух средних значений:

Статистическа
я
характеристика

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

Критерий :
проверки:

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



Равенство двух выборочных дисперсий:

Статистическая характеристика:

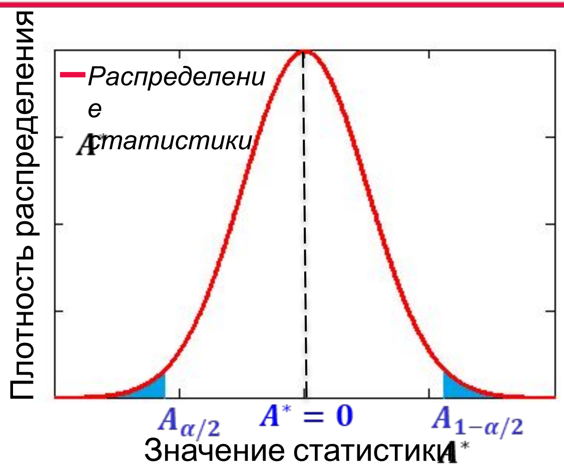
$$F_{NM} = \frac{s_1{}^2}{s_2{}^2}$$

Критерий проверки:

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

F - табулированное распределение

Критерии для асимметрии и эксцесса

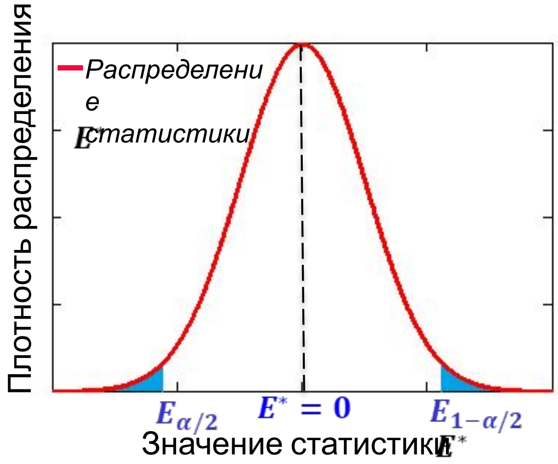


Суждение об асимметрии распределения

Базовая гипотеза: $A^* = 0$

Критерий проверки: $A^* \in \left(-\infty, A_{\frac{\alpha}{2}}\right) \cup \left(-\infty, A_{1-\frac{\alpha}{2}}\right)$

$$A \sim \mathcal{N}\left(0, \sqrt{\frac{6N(N-1)}{(N+3)(N+1)(N-2)}}\right)$$



Суждение о выраженности пика распределения

Базовая гипотеза: $E^* = 0$

Критерий проверки: $E^* \in \left(-\infty, E_{\frac{\alpha}{2}}\right) \cup \left(-\infty, E_{1-\frac{\alpha}{2}}\right)$

$$E \sim \mathcal{N}\left(0, \sqrt{\frac{24N(N-1)^2}{(N+5)(N+3)(N-2)(N-3)}}\right)$$

Метод Монте-Карло для проверки гипотез

Для базовой гипотезь $\mathrm{H}_0: \Theta^* = \Theta_0$ и альтернативь $\mathrm{H}_1: \Theta^* = \Theta_1$

- Задать аналитическую модель распределе $F_\xi(x, \Theta)$
- Выбрать метод моделирования случайной величины с этим распределением
- Сгенерировать M выборок по зако $F_\xi(x, \Theta_0)$,и M выборок по закону $F_\xi(x, \Theta_1)$
- Рассчитать по каждой выборке точечную оценку парамет Θ_1^* в Θ_0^* и
- Оценить распределени $F_{\Theta_1}(x)$ $F_{\Theta_0}(x)$ оценок параметров

Для заданного значения уровня значимос α найти ошибку второго рода γ .

Задача об отборе на выставку кошек (продолжение)

$Mx0 := 5.97$ $N := 30$

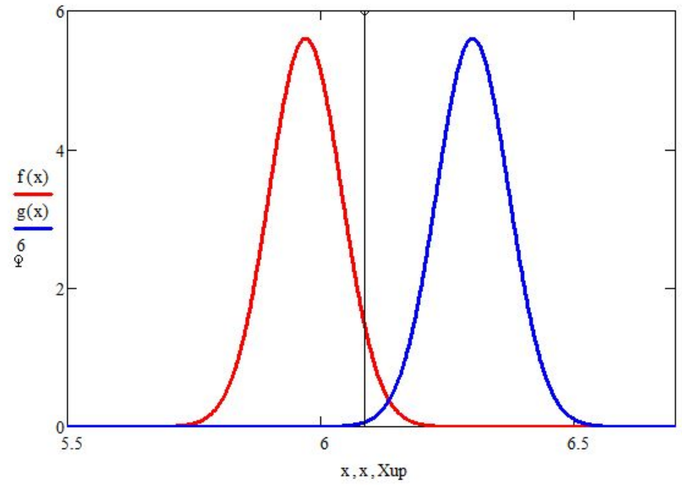
$Sx := 0.39$

$Mx1 := 6.3$

$x := 5.5, 5.501 .. 7.0$

$f(x) := \mathrm{dnorm}\left(x, Mx0, \frac{Sx}{\sqrt{N}}\right)$ $g(x) := \mathrm{dnorm}\left(x, Mx1, \frac{Sx}{\sqrt{N}}\right)$

$Xup := \mathrm{qnorm}\left(0.95, Mx0, \frac{Sx}{\sqrt{N}}\right)$ $\boxed{Xup = 6.087}$

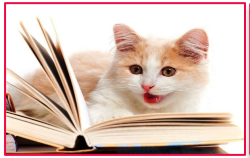


$G := \int_{5.5}^{Xup} g(x)\,dx$ $\boxed{G = 0.14\%}$

Средний вес по выборке из 30 кошек составил 5.97 кг, а СКО – 0.39 кг.

Можно ли с 10% уровнем значимости считать, что математическое ожидание веса кошек по генеральной совокупности составляет 6.0 кг?

Какова при этом вероятность реализации гипотезы, что математичес-



3.1

2.3

Развитие: статистические игры и решения

Решение $d_i \in D$ и его *функция потерь* $L(x,d)$

Цель: выбор оптимальной последовательности решений

$$d_1 \to d_2 \to \cdots \to d_m$$

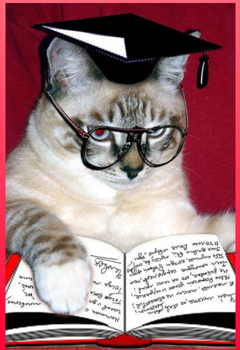
на основе ограниченного числа наблюдений $(x_1, \ldots, x_N)$
за случайной величиной Ξ

Риск: средние потери при решении d_k

$$R_{d_k} = \int_{-\infty}^{+\infty} L(x, d_k) f_{\Xi}(x)\,dx$$

Смысл игры: предугадать реакцию «противника» (значение Ξ) так, чтобы минимизировать свои риски

Автор признателен всем котикам (и хозяевам) за возможность некоммерческого использования их изображений, размещенных в публичном Интернете



Александр Валерьевич Бухановский, Анна Владимировна Калюжная
boukhanovsky@mail.ifmo.ru

ITsMOre than a UNIVERSITY

УНИВЕРСИТЕТ ИТМО

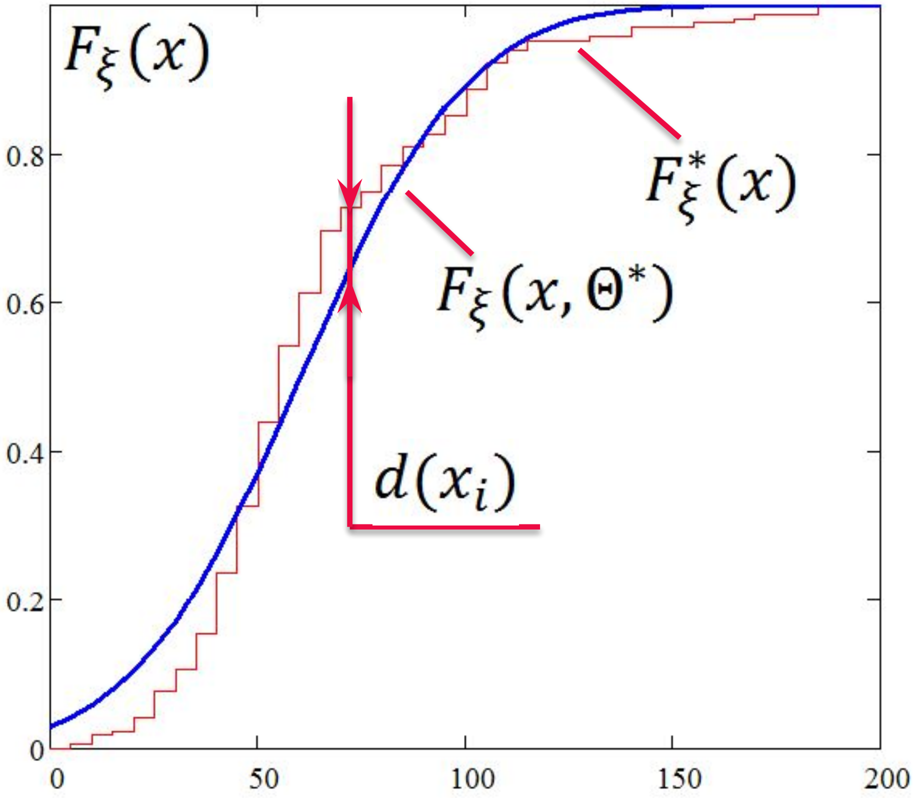
ЗАНЯТИЕ 3

ТЕОРИЯ И ПРАКТИКА СТАТИСТИЧЕСКИХ ВЫВОДОВ

Раздел 4. Конструктивные модели распределений

Александр Валерьевич Бухановский, Анна Владимировна Калюжная

Подбор распределения как задача приближения



Минимизация нормы невязки: $\|d(x)\|_{x\in[a,b]} \underset{\Theta}{\rightarrow} min$

Идея метода наименьших квадратов (МНК)

УНИВЕРСИТЕТ ИТМО

Невязка как ***евклидово расстояние***:

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

Идея метода (минимизация суммы квадратов невязок по *L* характерным квантилям):

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

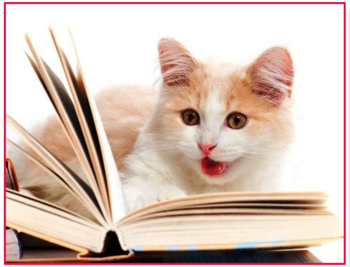
Реализация метода (система уравнений):

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

МНК для линейно-масштабируемых величин

Квантиль линейно-масштабируемой величины:

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



1.1

$$U_{p_k} = F^{-1}_{\xi(0)}(p_k)$$ - квантиль нормированного модельного распределения

Реализация МНК: система из двух линейных алгебраических уравнений

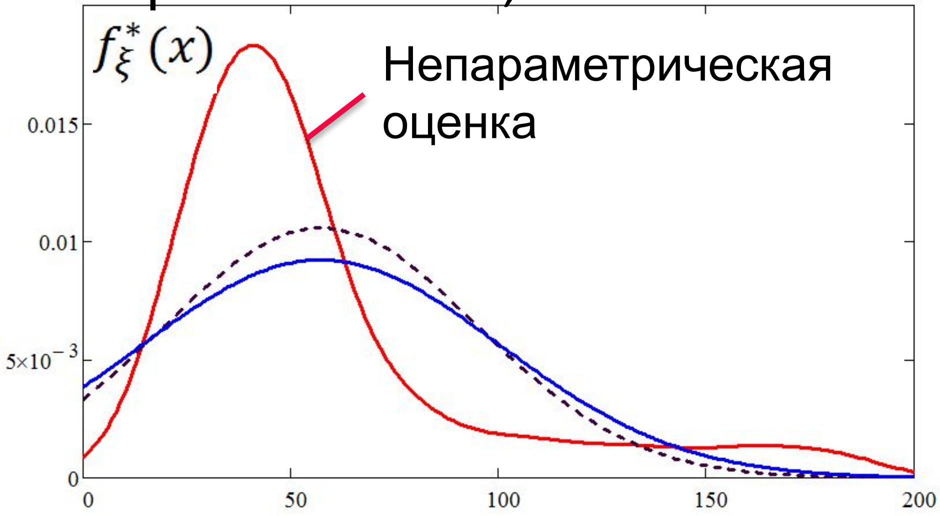
$$\begin{cases} L\theta_1 + \theta_2 \sum_{k=1}^{L} U_{p_k} = \sum_{k=1}^{L} x^*_{p_k} \\ \theta_1 \sum_{k=1}^{L} U_{p_k} + \theta_2 \sum_{k=1}^{L} U^2_{p_k} = \sum_{k=1}^{L} U_{p_k} x^*_{p_k} \end{cases}$$

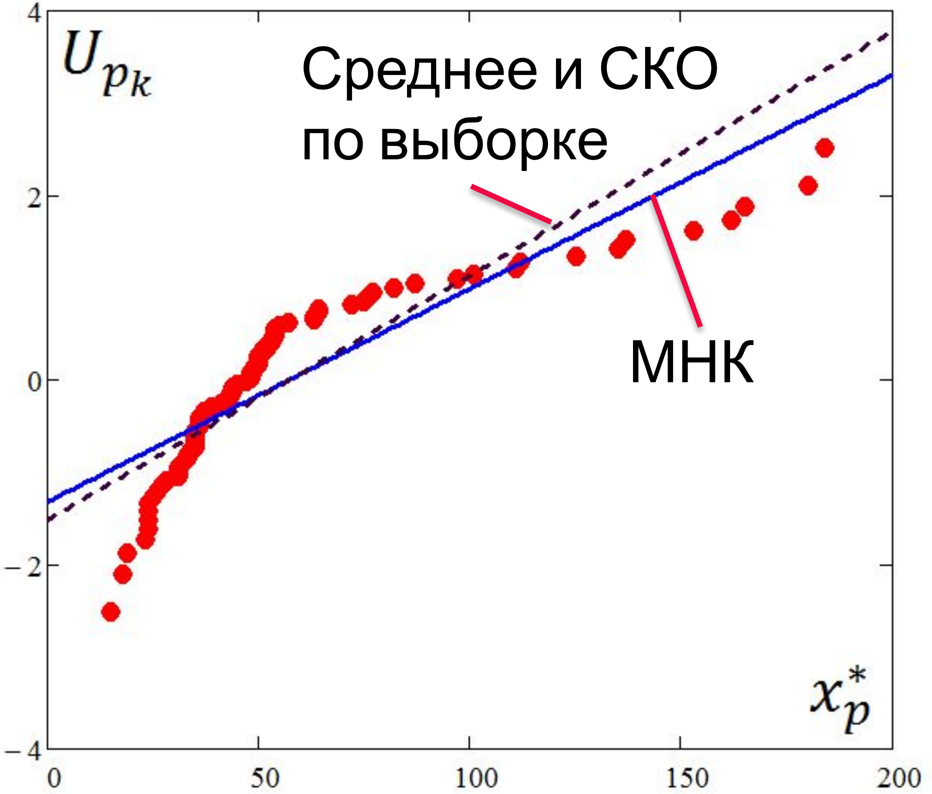
Простейшее приближение: МНК по всей выборке

Задача: приблизить модельное распределение к выборке, которая ***не вполне*** ему соответствует

Все выборочные квантили равнозначны,

$$L = N$$



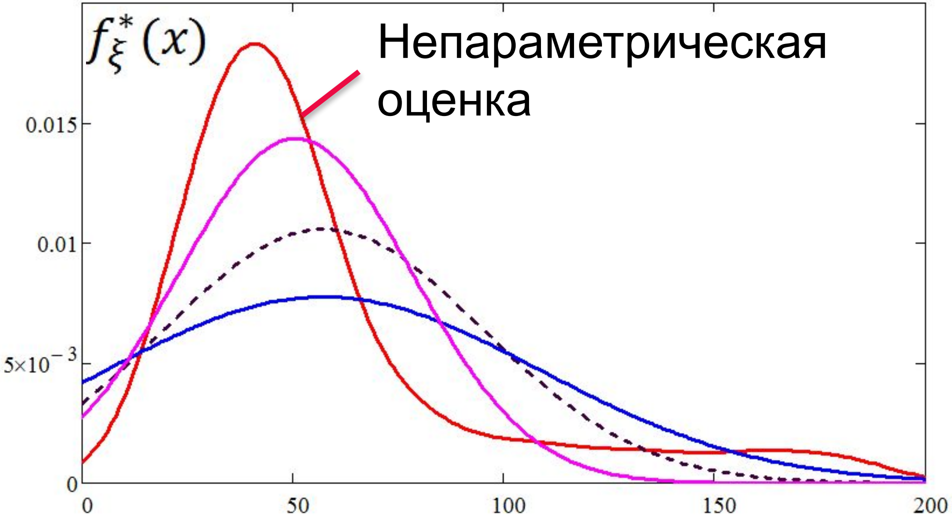


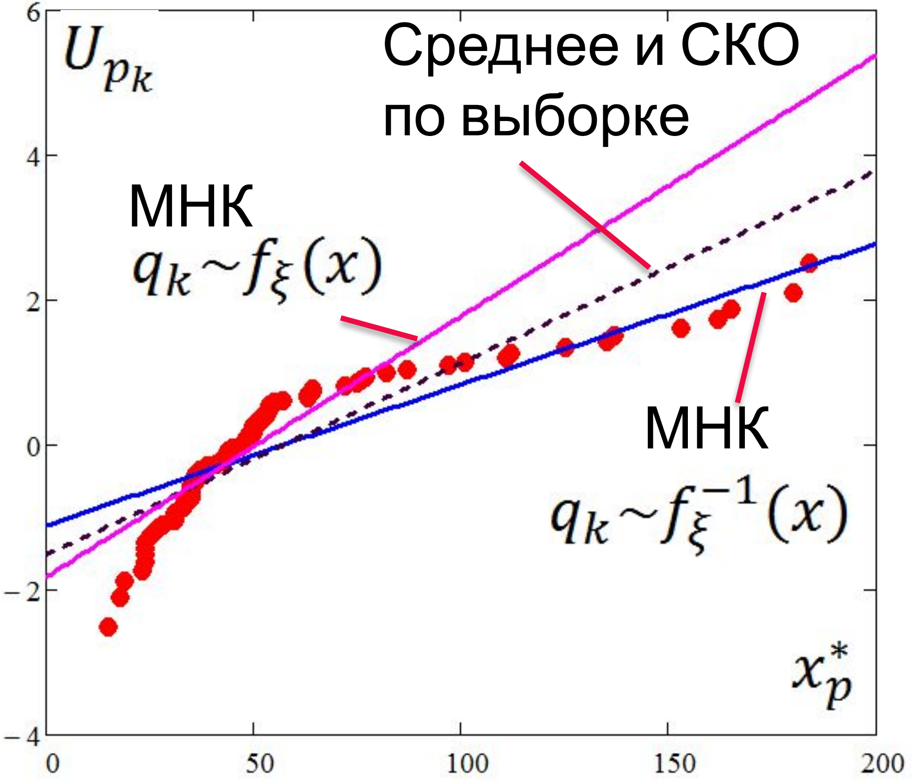
Взвешенный МНК: учитываем особенности

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

Целевая функция с весам q_k

Выборочные квантили **не**равнозначны, $L = N$



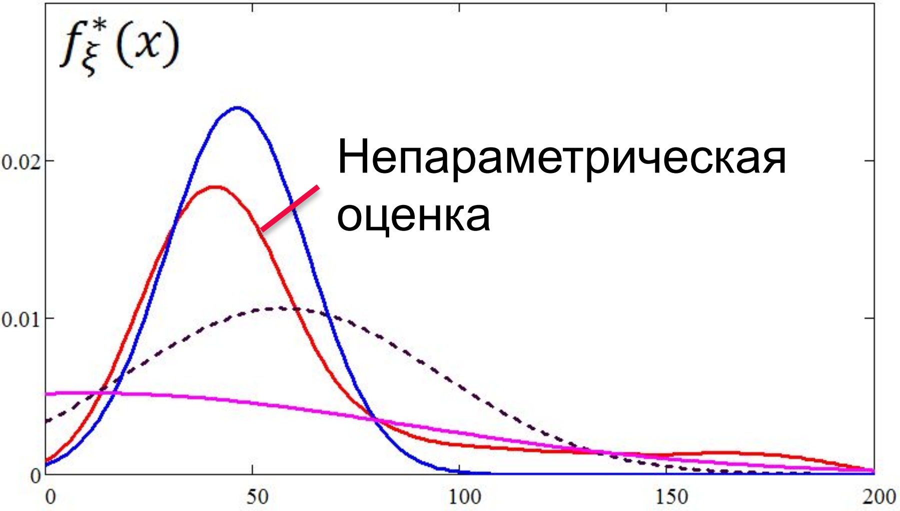


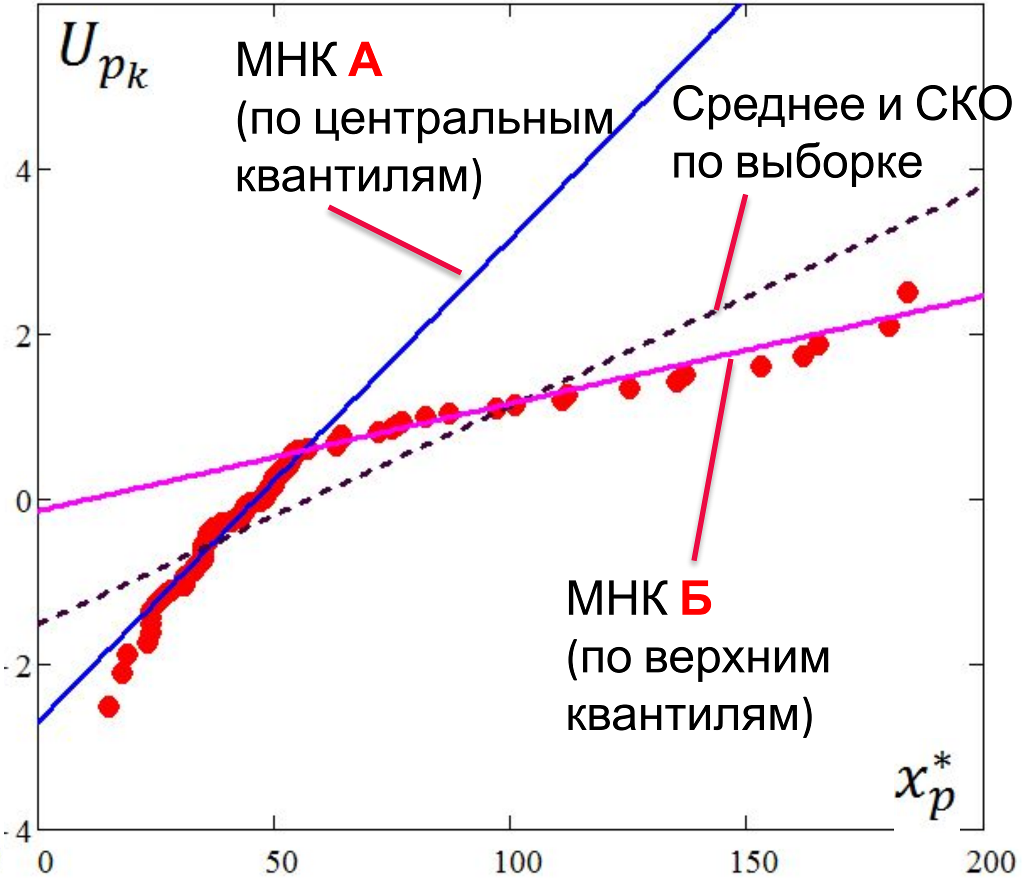
МНК по избранным квантилям

Предположение: избранные квантили равнозначны, $L < N$

А: квантили (10%,25%,35%,50%,65%,75%)

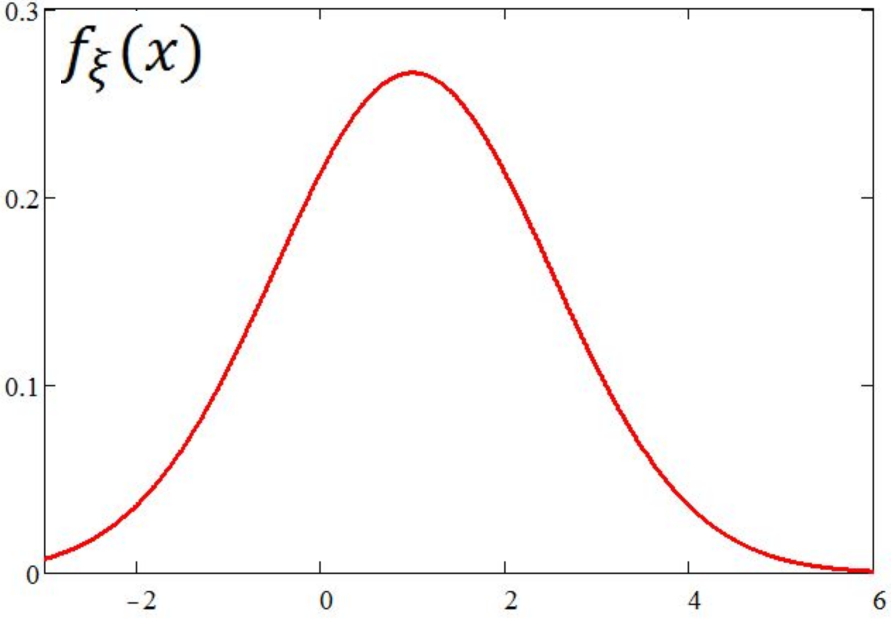
Б: квантили (75%,85%,90%,95%, 97%,99%)

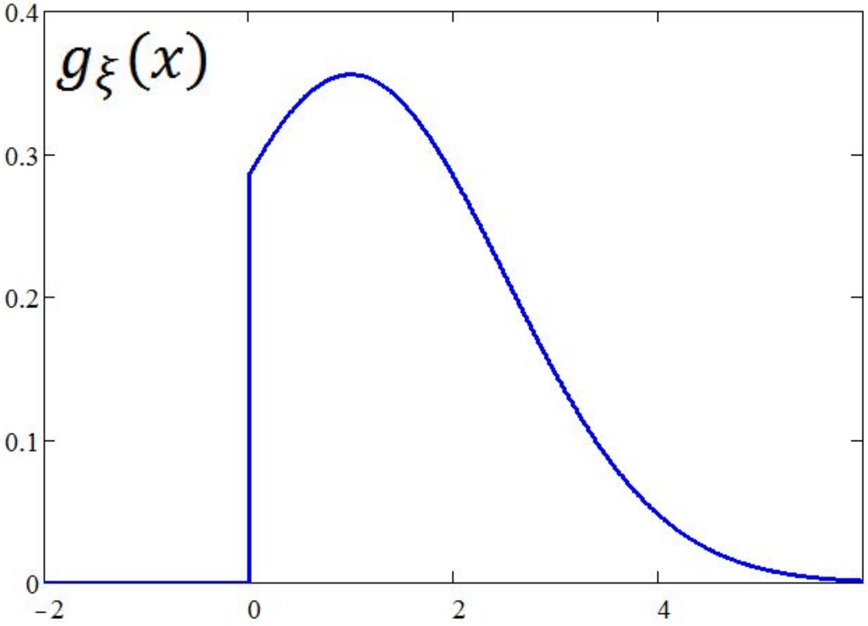




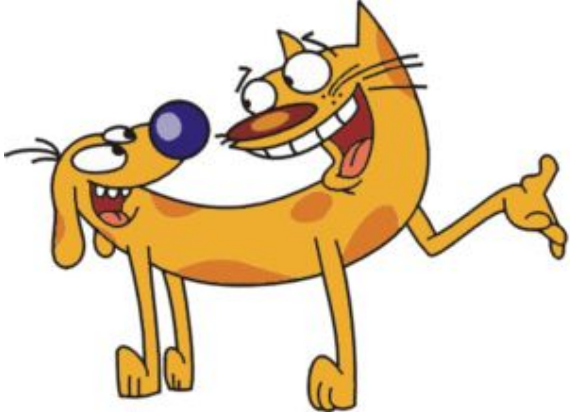
Модель усеченного распределения

$$g_\xi(x) = \begin{cases} f_\xi(x)/C, x \geq a \\ 0, x < a \end{cases} \qquad C = \int_a^{+\infty} f_\xi(x)dx$$



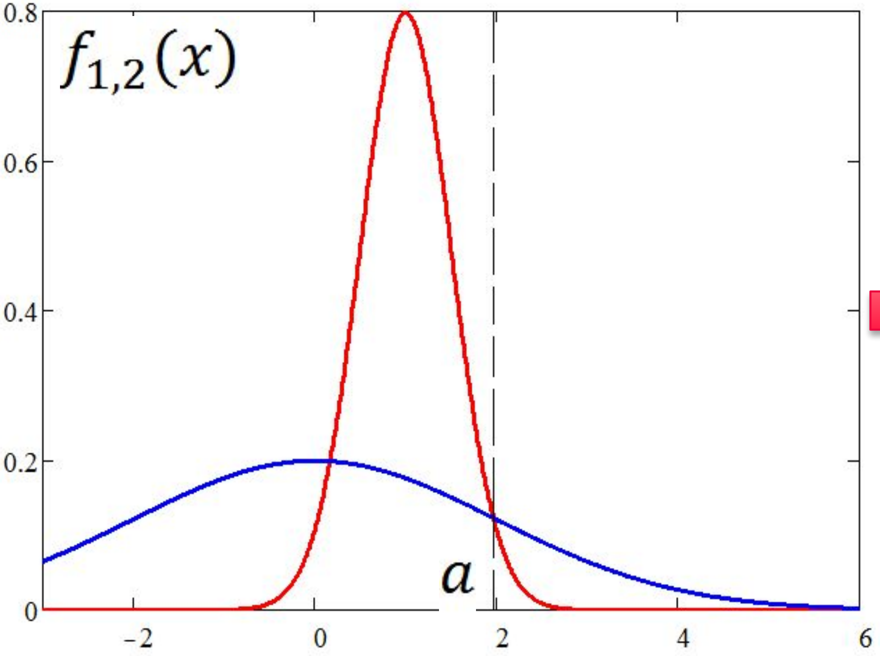


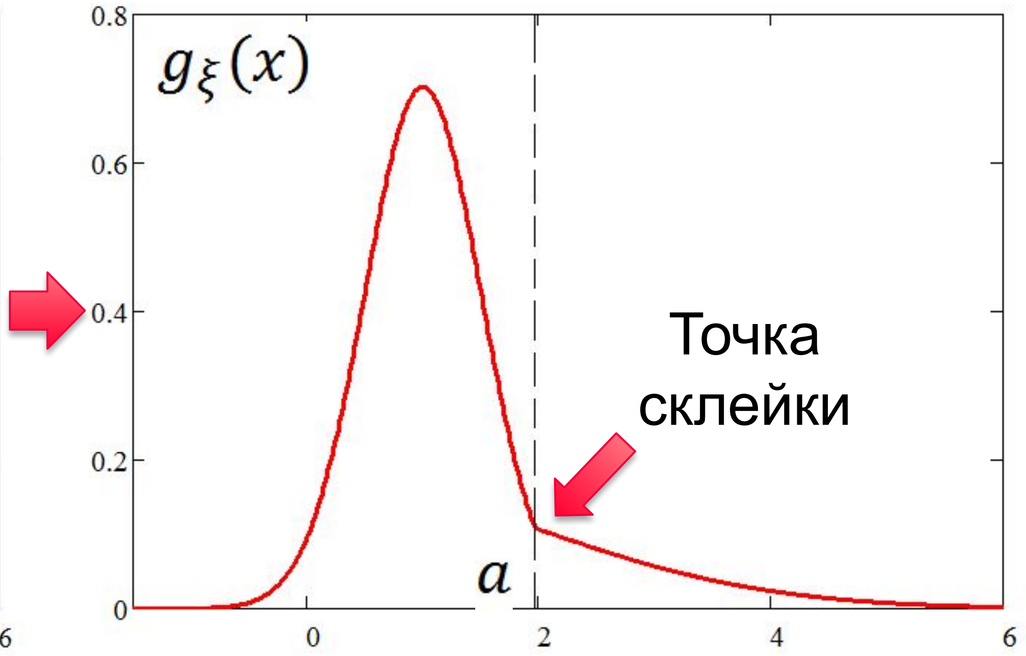
Модель склейки распределений



$$g_\xi(x) = \begin{cases} f_2(x)/C, x \geq a \\ f_1(x)/C, x < a \end{cases}$$

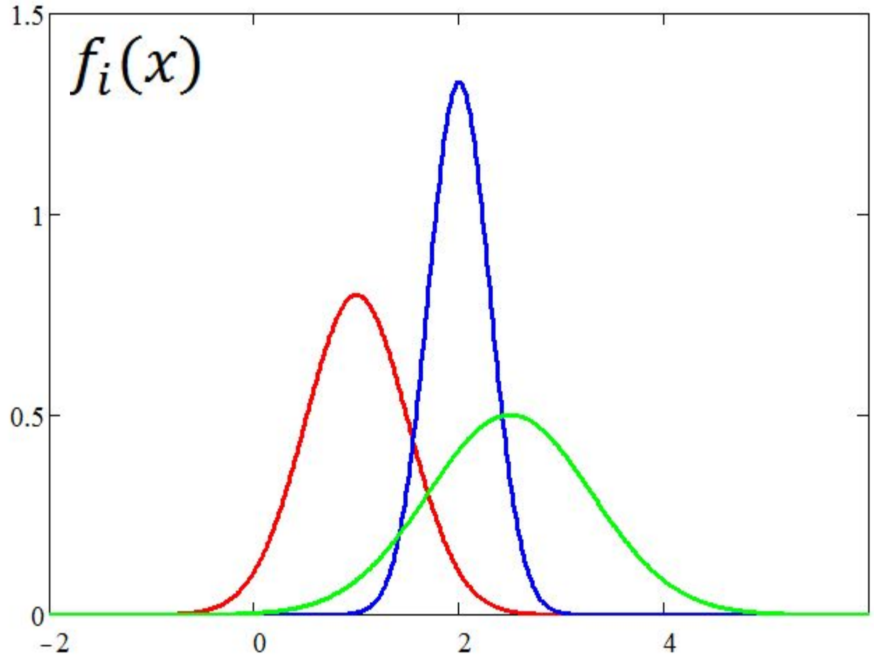
C – нормировочная константа

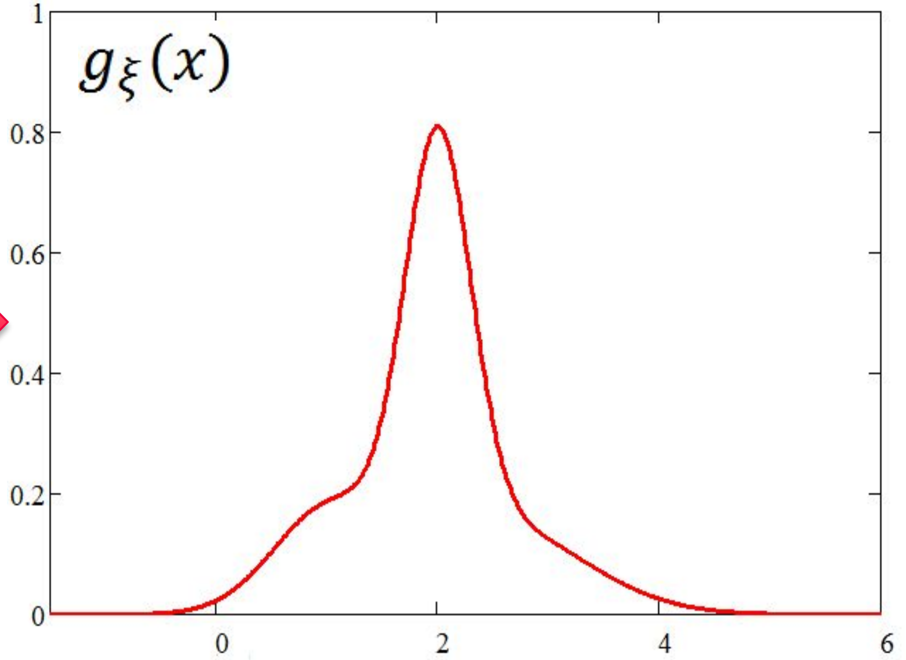




Модель смеси распределений

$$g_\xi(x) = \sum_{i=1}^{Q} p_i f_i(x) \qquad \sum_{i=1}^{Q} p_i = 1$$

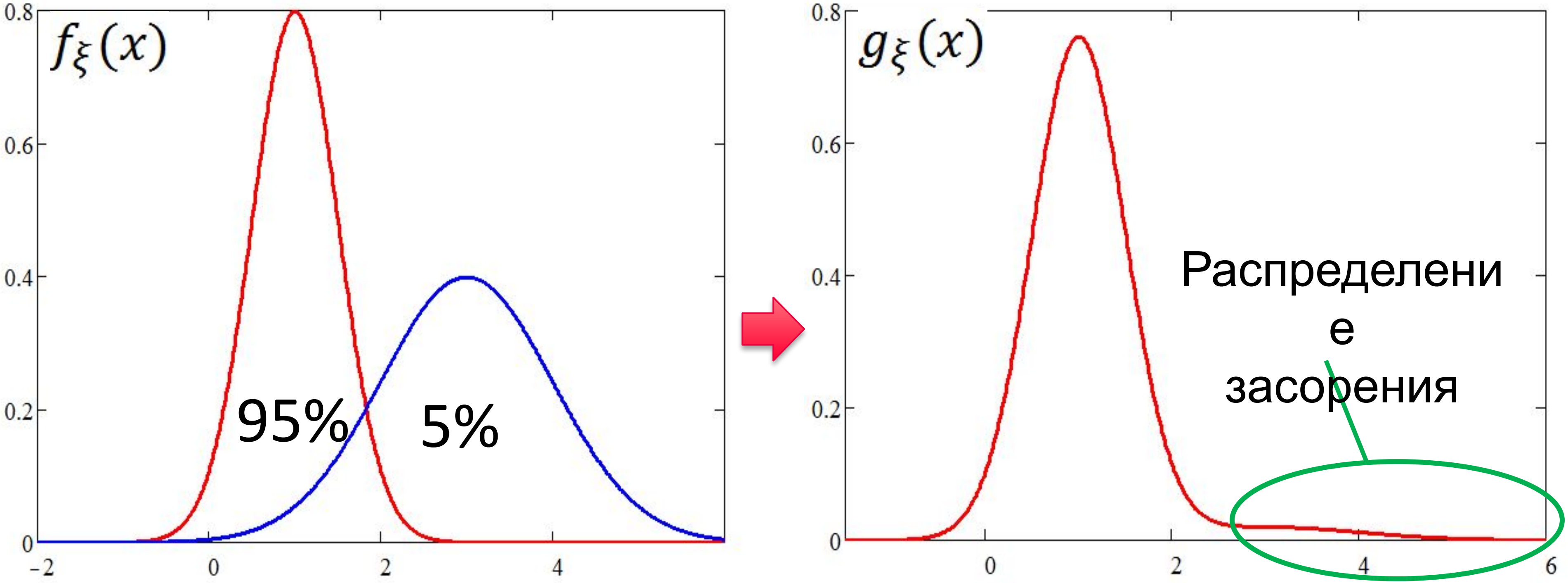




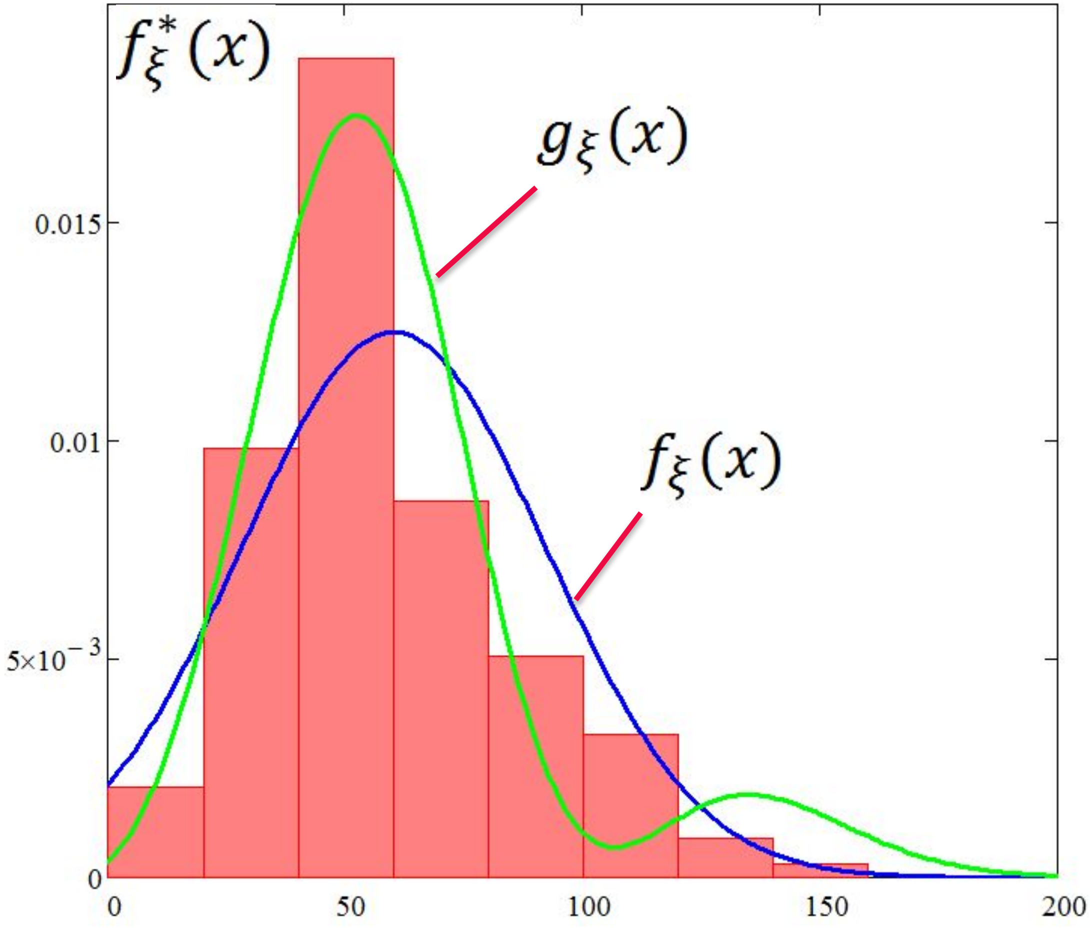
Модель распределения с засорением

$$g_\xi(x) = (1-\varepsilon)f_\xi(x,\Theta) + \varepsilon f_\xi(x, m\Theta)$$

m - параметр масштаба засорения



Модель в форме ряда Эджворта



Разложение в окрестности модельного р

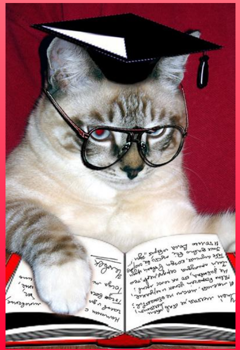
$$g_\xi(x) = f_\xi(x)\sum_{j=1}^{n} a_j \varphi_j(x)$$

$$a_j = f_j(m_1, \dots m_n)$$

Обобщение: конструктивные распределения

- **Метод МНК:** удачный механизм для «натягивания» моделей распределений на реальные данные
- **Механизмы управления детализацией модели распределения:** выбор весов и расположения квантилей для МНК.
- **Составные модели распределений** (усечение, склейка): применение МНК к разным частям вариационного ряда
- **Модели на основе смесей:** просто использовать, сложно оценивать
- **Разложения в ряд Эджворта:** эффективно, но лишь при малых отклонениях

Автор признателен всем котикам (и хозяевам) за возможность некоммерческого использования их изображений, размещенных в публичном Интернете



Александр Валерьевич Бухановский
boukhanovsky@mail.ifmo.ru

ITsMOre than a UNIVERSITY

УНИВЕРСИТЕТ ИТМО

ЗАНЯТИЕ 3

ТЕОРИЯ И ПРАКТИКА СТАТИСТИЧЕСКИХ ВЫВОДОВ

Раздел 5. Практическая реализация с использованием инструментов компьютерной математики

Александр Валерьевич Бухановский, Анна Владимировна Калюжная

Инструменты компьютерной математики

1. MATLAB
2. R
3. Python
4. MS Excel
5. Statistica

Интервальное оценивание на основе Монте-Карло в MATLAB

PUBLISH VIEW

Breakpoints | Run | Run and Advance | Run Section | Advance | Run and Time

BREAKPOINTS | RUN

MATLAB

Editor - C:\Users\irina\Documents\ITMO\MATLAB\Monte_

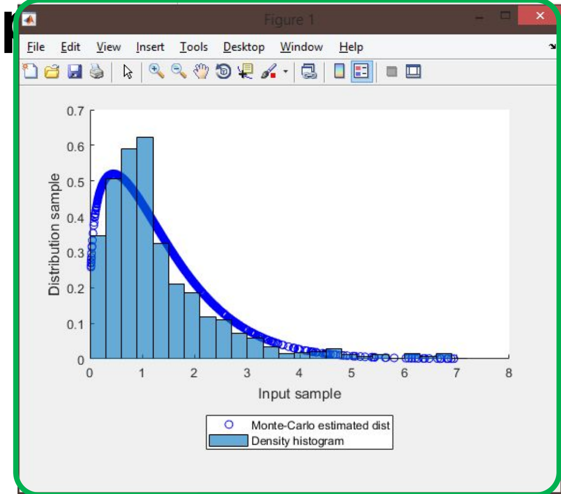
Monte_Carlo.m | The_first_example.m

```
clear,clc,close all
%% Open the file
waves = dlmread('dataset.txt');
p = mle('weib',waves);
pd = wblpdf(waves,p(1),p(2));
a=0;
b=max(waves);
%% Geometric method for generating samples
N = 100;
M = 100;
X = zeros(N, M);
for i = 1:N
    for j = 1:M
        y1=rand;
        y2=rand;
        y1m=a+(b-a)*y1;
        y2m=max(pd)*y2;
    while y2m>wblpdf(y1m,p(1),p(2))
       y1=rand;
        y2=rand;
        y1m=a+(b-a)*y1;
        y2m=max(pd)*y2;
    end
    X(i,j) = y1m;
    end
end
```

```
%% Point estimation of lambda and k
point_est = zeros (M,2);
for i=1:M
    phat = mle('weib',X(:,i)');
    point_est(i,1) = phat(1);
    point_est(i,2) = phat(2);
end
%% Interval estimation by quantiles 2.5% and 97.5%
lambda_q1 = quantile(point_est(:,1)',0.025);
lambda_q2 = quantile(point_est(:,1)',0.975);
k_q1 = quantile(point_est(:,2)',0.025);
k_q2 = quantile(point_est(:,2)',0.975);
l = mean(point_est(:,1));
k = mean(point_est(:,2));
array_k = [k_q1, k, k_q2];
array_l = [lambda_q1,l,lambda_q2];
Y = wblpdf(waves,l,k);
```

```
figure(1)
hold on
plot(waves,Y,'bo')
h = histogram(waves,'Normalization','pdf');
ylabel('Distribution sample')
xlabel('Input sample')
legend({'Monte-Carlo estimated dist','Density histogram'},'Location', 'southoutside');
```

```
%% Tabel of results
T1 = array2table(array_l, 'VariableNames',{'lambda1','Lambda','Lambda2'});
display(T1);
T2 = array2table(array_k, 'VariableNames',{'k1','k','k2'});
display(T2);
```



Command Window

New to MATLAB? See resources for Getting Started.

T1 =

1×3 table

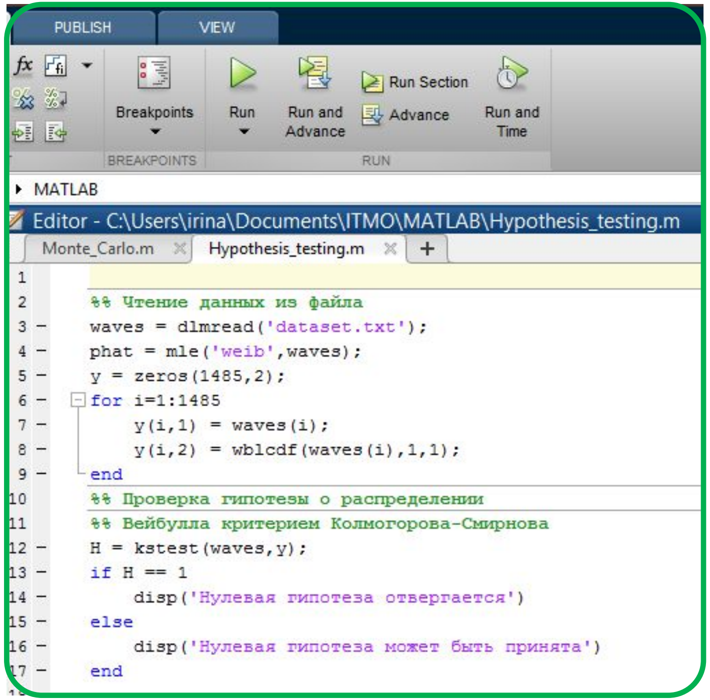
lambda1	Lambda	Lambda2
1.2157	1.4161	1.6441

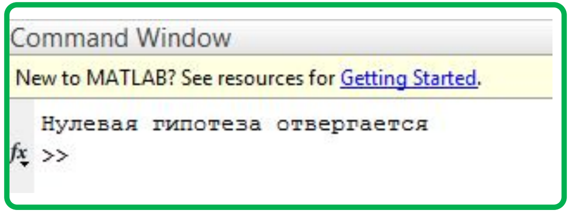
T2 =

1×3 table

k1	k	k2
1.1044	1.2924	1.5805

Тест Колмогорова-Смирнова в MATLAB

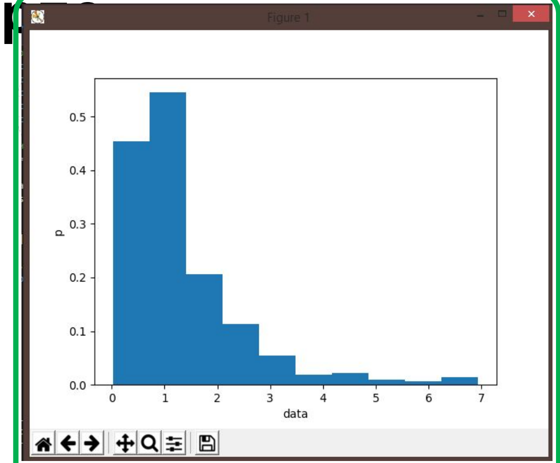


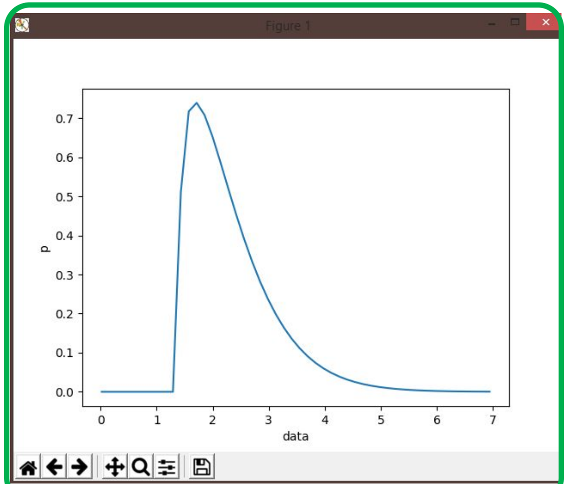


Интервальное оценивание на основе Монте-Кар
в Python

```
File  Edit  View  Project  Build  Debug  Team  Tools  Test  Analyze  Win
Debug  Any CPU  Attach
fit_distr.py   Lecture_3.py*

# Импорт пакетов
import matplotlib.pyplot as plt
import numpy as np
import scipy.stats
import seaborn as sns
import random

# Чтение значений случайной величины из файла
data = np.loadtxt('dataset.txt')
a = 0
b = max(data)
params = scipy.stats.weibull_min.fit(data)
pdf = scipy.stats.weibull_min.pdf(data, *params)
cdf = scipy.stats.weibull_min.cdf(data, *params)
X = []
for i in range(100):
    X.append([])
    for j in range(100):
        y1 = random.random()
        y2 = random.random()
        y1m = a + (b-a)*y1
        y2m = max(pdf)*y2
        while y2m > scipy.stats.weibull_min.pdf(y1m,*params):
            y1 = random.random()
            y2 = random.random()
            y1m = a + (b-a)*y1
            y2m = max(pdf)*y2
        X[i].append(y1m)
```

```
ki = []
li = []
for i in range(100):
    params = scipy.stats.weibull_min.fit(X[i:])
    ki.append(params[0])
    li.append(params[2])
k1 = np.quantile(ki,0.025)
k2 = np.quantile(ki,0.975)
l1 = np.quantile(li,0.025)
l2 = np.quantile(li,0.975)
k = np.mean(ki)
l = np.mean(li)
print("k param")
print(k1," ",k," ",k2)
print("lambda param")
print(l1," ",l," ",l2)
plt.hist(sorted(data), 10,normed=1)
plt.xlabel('data')
plt.ylabel('p')
pdf = scipy.stats.weibull_min.pdf(x, *(k,l))
plt.plot(x, pdf)
plt.show()
```






```
C:\Program Files (x86)\Microsoft Visual Studio\Shared\Python36_64\python.e
k param
1.2438805598748481   1.2568268668234348   1.3141535399463202
lambda param
1.3547662100147817   1.4001340128978037   1.4230322081626214
```

Тест Колмогорова-Смирнова в Python

УНИВЕРСИТЕТ ИТМО

File Edit View Project Build Debug Team Tools Test Analyze

Debug Any CPU

fit_distr.py Lecture_3.py*

```python
# Импорт пакетов
import matplotlib.pyplot as plt
import numpy as np
import scipy.stats
import seaborn as sns
import random

# Чтение значений случайной величины из файла
data = np.loadtxt('dataset.txt')
```

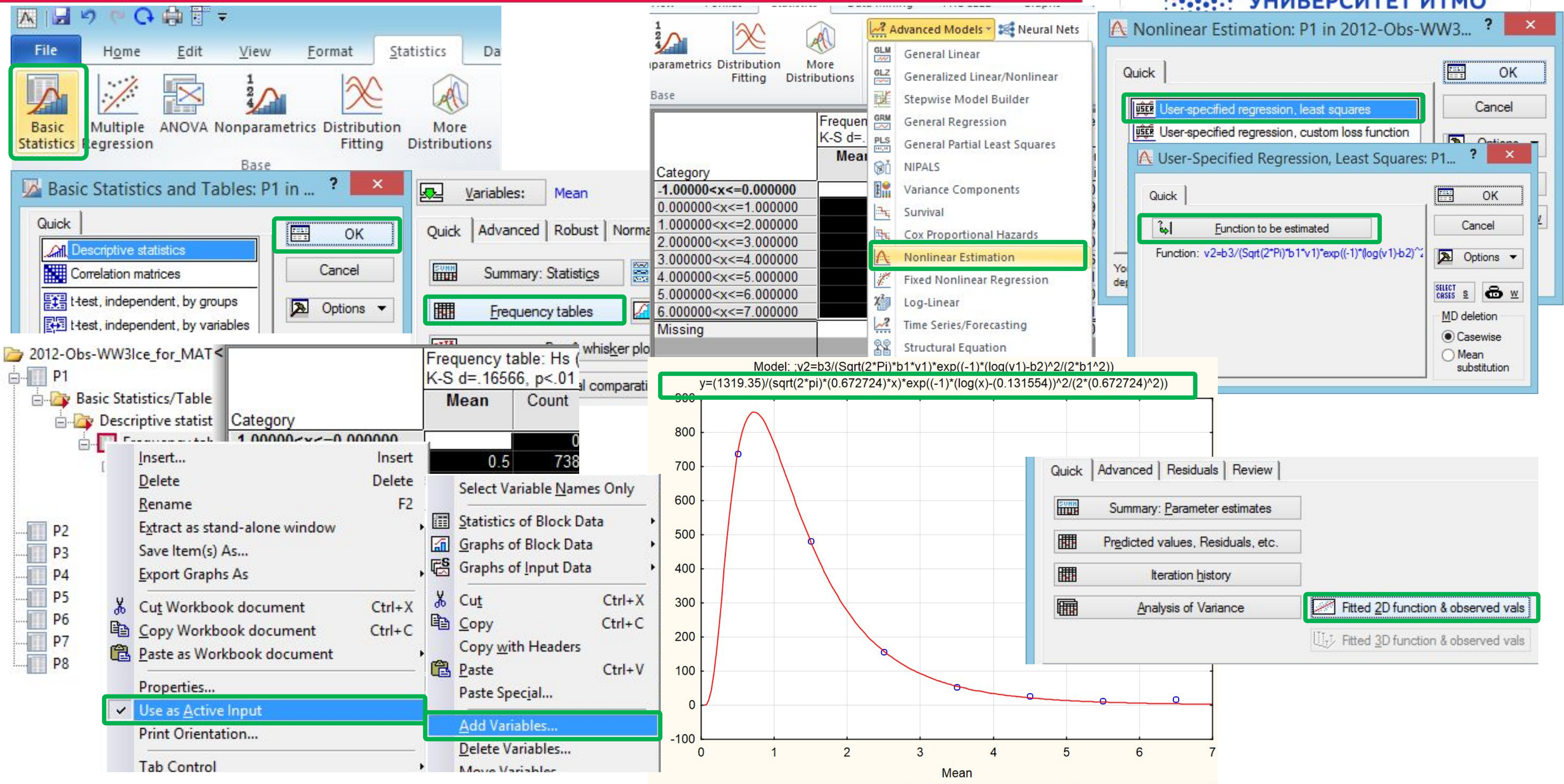
```python
k = np.mean(ki)
l = np.mean(li)
print("k param")
print(k1," ",k," ",k2)
print("lambda param")
print(l1," ",l," ",l2)

x = np.linspace(np.min(data), np.max(data))
print(scipy.stats.kstest(data,'norm'))
```

C:\Program Files (x86)\Microsoft Visual Studio\Shared\Python36_64\python.exe

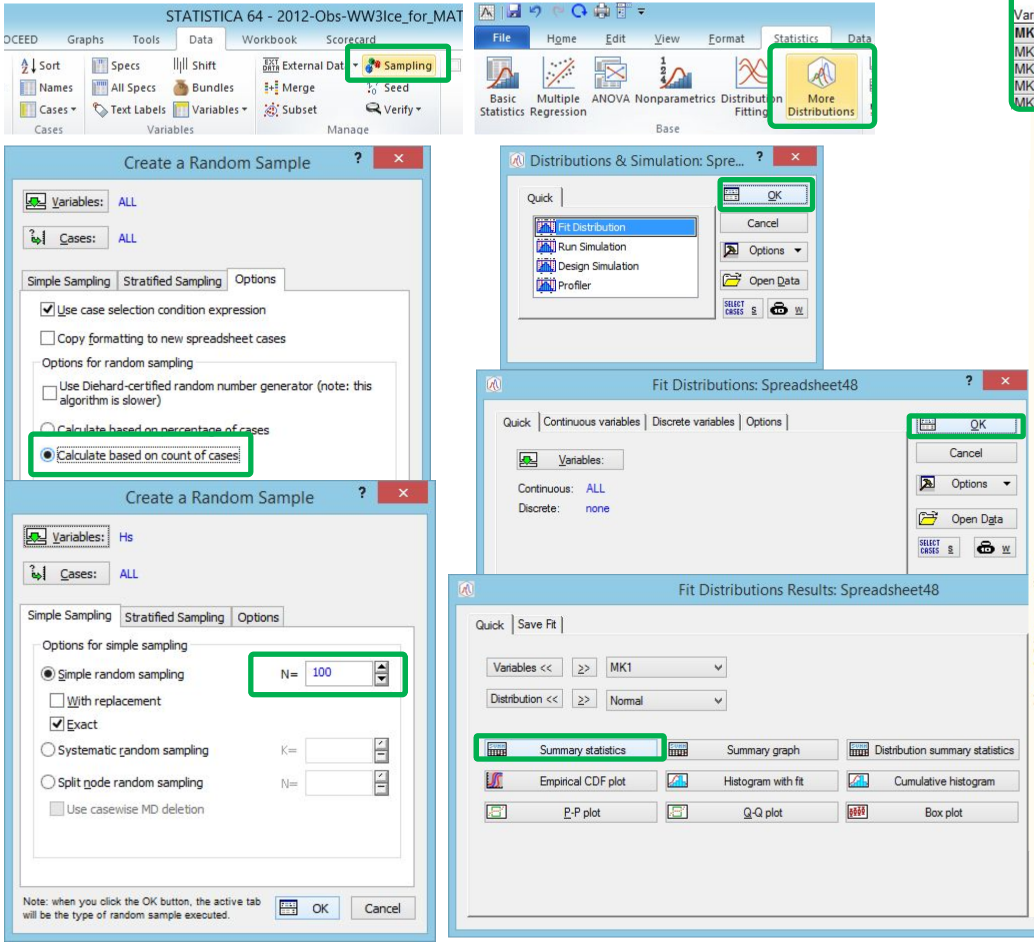
```
k param
1.2255553775709278    1.2450224347644534    1.2704436286733507
lambda param
1.3778728553729875    1.3933293279795977    1.4307822839498954
KstestResult(statistic=0.535224978668905, pvalue=0.0)
Press any key to continue . . .
```

ITMOre than a UNIVERSITY

Метод наименьших квадратов в Statistica(1/2)

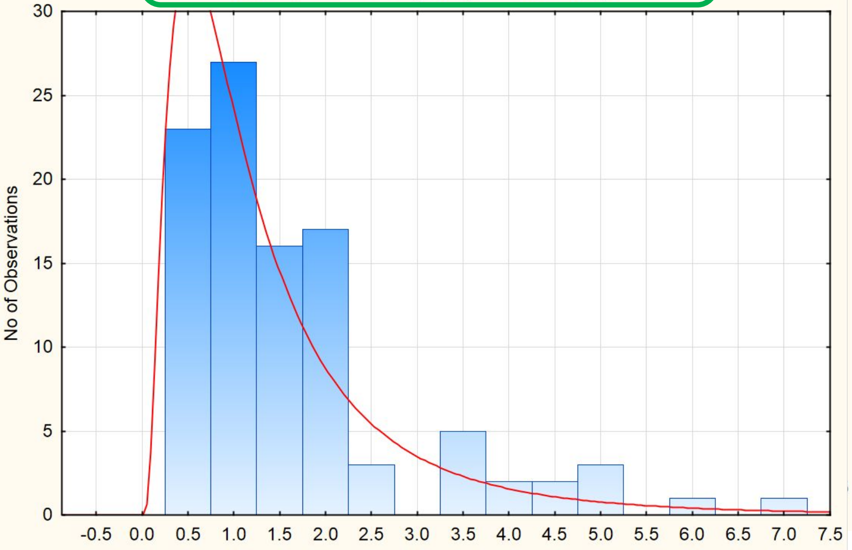


Метод Монте-Карло в Statistica(2/2)

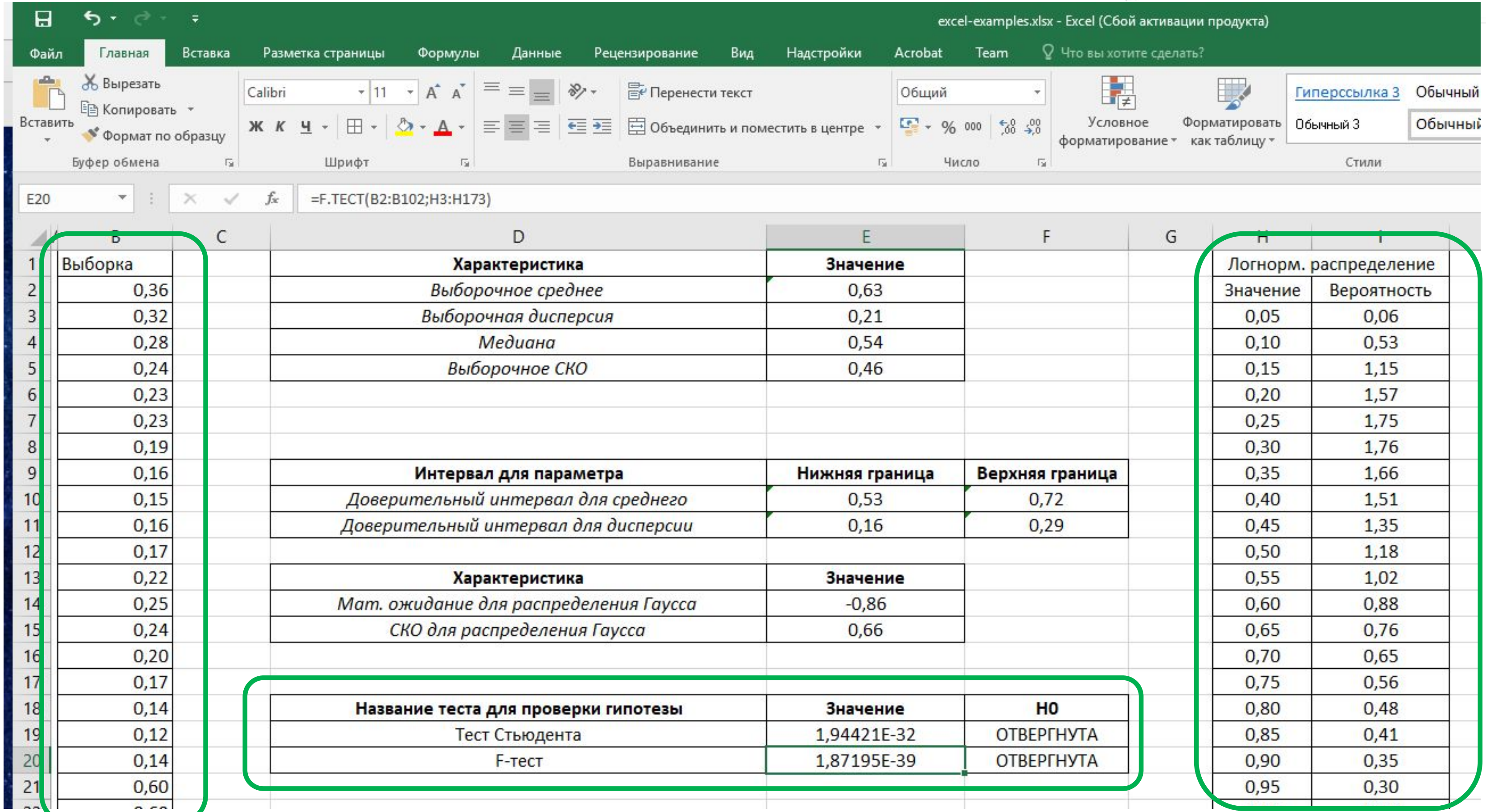


Variable	Mean	Confidence -95.000%	Confidence 95.000%	Lower Quartile	Upper Quartile	Std.Dev.	Confidence SD -95.000%	Confidence SD +95.000%
MK1	1.324220	1.090281	1.558159	0.544500	1.731500	1.179000	1.035170	1.369615
MK2	1.149380	0.962073	1.336687	0.593000	1.377500	0.943984	0.828825	1.096603
MK3	1.324220	1.090281	1.558159	0.544500	1.731500	1.179000	1.035170	1.369615
MK4	1.554220	1.285686	1.822754	0.722500	1.911500	1.353349	1.188250	1.572152
MK5	1.439780	1.184912	1.694648	0.579000	1.779500	1.284479	1.127781	1.492147

Histogram of MK1
MK1 = 100*0.5000*Lognorm(x,-0.0710,0.8777)
Histogram of MK2
MK2 = 100*0.5000*Lognorm(x,-0.1518,0.8667)
Histogram of MK3
MK3 = 100*0.5000*Lognorm(x,-0.0797,0.9200)
Histogram of MK4
MK4 = 100*0.5000*Lognorm(x,0.0859,0.9214)
Histogram of MK5
MK5 = 100*0.5000*Lognorm(x,0.0295,0.8258)



Проверка статистических гипотез в Excel (1/1)



	B
1	Выборка
2	0,36
3	0,32
4	0,28
5	0,24
6	0,23
7	0,23
8	0,19
9	0,16
10	0,15
11	0,16
12	0,17
13	0,22
14	0,25
15	0,24
16	0,20
17	0,17
18	0,14
19	0,12
20	0,14
21	0,60

Характеристика	Значение
Выборочное среднее	0,63
Выборочная дисперсия	0,21
Медиана	0,54
Выборочное СКО	0,46

Интервал для параметра	Нижняя граница	Верхняя граница
Доверительный интервал для среднего	0,53	0,72
Доверительный интервал для дисперсии	0,16	0,29

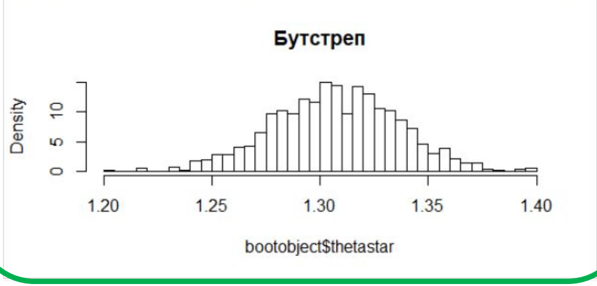
Характеристика	Значение
Мат. ожидание для распределения Гаусса	-0,86
СКО для распределения Гаусса	0,66

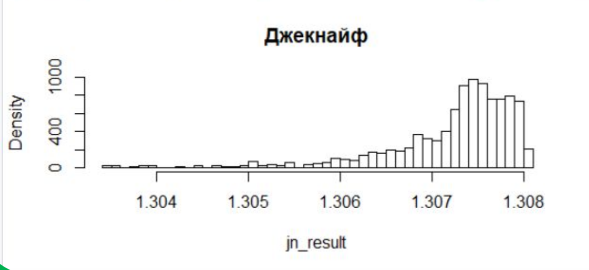
Название теста для проверки гипотезы	Значение	H0
Тест Стьюдента	1,94421E-32	ОТВЕРГНУТА
F-тест	1,87195E-39	ОТВЕРГНУТА

Логнорм. распределение	
Значение	Вероятность
0,05	0,06
0,10	0,53
0,15	1,15
0,20	1,57
0,25	1,75
0,30	1,76
0,35	1,66
0,40	1,51
0,45	1,35
0,50	1,18
0,55	1,02
0,60	0,88
0,65	0,76
0,70	0,65
0,75	0,56
0,80	0,48
0,85	0,41
0,90	0,35
0,95	0,30

Интервальное оценивание на основе Монте-Карло в R (1/2)

```r

library(bootstrap)
library(fitdistrplus)
library(goftest)

#Чтение значений случайное величины
real_sample=read.table("X.csv",sep=";",header=TRUE)$x

#функция вычисления среднего по выборке
meanfun <- function(data, indices){
  d <- data[indices]
  return(mean(d))
}

#Моделирование распределения оценки мат. ожидания методом бутстреп
br_result = bootstrap(x=real_sample,nboot=1000,theta=meanfun)$thetastar
hist(br_result,breaks = 50,freq = FALSE,main="")
title("Бутстреп")
pint=quantile(br_result,c(0.025,0.975))
c(pint[1],mean(br_result),pint[2])

#Моделирование распределения оценки мат. ожидания методом джекнайф
jn_meanfun <- function(d){
  return(mean(d))
}
jn_result=jackknife(real_sample,jn_meanfun)$jack.values
hist(jn_result,breaks = 50,freq = FALSE,main="")
title("Джекнайф")
pint=quantile(jn_result,c(0.025,0.975))
c(pint[1],mean(jn_result),pint[2])
```






```
> c(pint[1],mean(br_result),pint[2])
    2.5%              97.5%
1.252100 1.307469 1.368992
> c(pint[1],mean(jn_result),pint[2])
    2.5%              97.5%
1.252100 1.307207 1.368992
>
```

Проверка статистических гипотез в R (2/2)

УНИВЕРСИТЕТ ИТМО

stat3.R

Source on Save Run

```r
#функция оценки параметров для логнормального распределения на основе метода максимального правдоподобия
theor_dist_fit_mle=fitdist(data=real_sample,distr="lnorm",method="mle")
#на основе метода квантилей
theor_dist_fit_qme=qmedist(data=real_sample,distr="lnorm",probs=c(0.25, 0.75))

#Проверка гипотезы методом Крамера-Мизеса
cvm_res=cvm.test(real_sample,"plnorm",mean=theor_dist_fit_mle$estimate[1],
                 sd=theor_dist_fit_mle$estimate[2])
cvm_res

#Проверка гипотезы методом Андерсена
ad_res=ad.test(real_sample,"plnorm",mean=theor_dist_fit_mle$estimate[1],
               sd=theor_dist_fit_mle$estimate[2])
ad_res

#Проверка гипотезы критерием Колмоговора-Смирнова
ks_res=ks.test(real_sample,"plnorm",mean=theor_dist_fit_mle$estimate[1],
               sd=theor_dist_fit_mle$estimate[2])
ks_res$statistic
```

D:/Projects/Lectures/ЦГ/Статистика-3/

```
> cvm_res

	Cramer-von Mises test of goodness-of-fit
	Null hypothesis: log-normal distribution

data:  real_sample
omega2 = 1.526, p-value = 0.0001498
```

```
> ad_res

	Anderson-Darling test of goodness-of-fit
	Null hypothesis: log-normal distribution

data:  real_sample
An = 8.1689, p-value = 9.216e-05
```

```
> ks_res

	One-sample Kolmogorov-Smirnov test

data:  real_sample
D = 0.068686, p-value = 1.644e-06
alternative hypothesis: two-sided
```

Console Terminal

D:/Projects/Lectures/ЦГ/Статистика-3/

```
> if (cvm_res$statistic>=qCvM(c(0.95))) {
+   print("Гипотеза H0 отклонена")
+ }
[1] "Гипотеза H0 отклонена"

> print(paste("Омега-квадрат =",round(ad_res$statistic,2),"95% квантиль распределения =",round(qAD(c(0.95)),2)))
[1] "Омега-квадрат = 8.17 95% квантиль распределения = 2.49"

> if (ad_res$statistic>=qAD(c(0.95))) {
+   print("Гипотеза H0 отклонена")
+ }
[1] "Гипотеза H0 отклонена"

> print(paste("p-value =",round(ks_res$p.value,6)))
[1] "p-value = 2e-06"

> if (ks_res$p.value<0.05) {
+   print("Гипотеза H0 отклонена")
+ }
[1] "Гипотеза H0 отклонена"
```