Распределения и статистики

Рябенко Евгений
riabenko.e@gmail.com

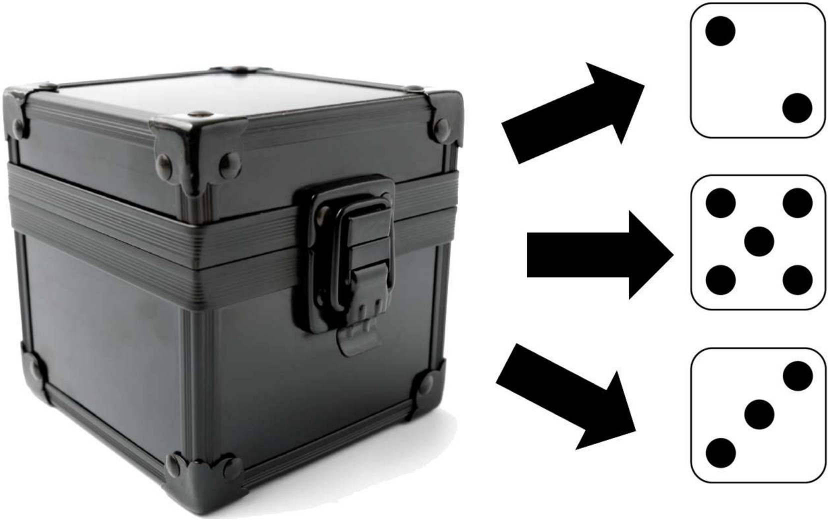
7 ноября 2016 г.

Случайность

Случайность

Случайность

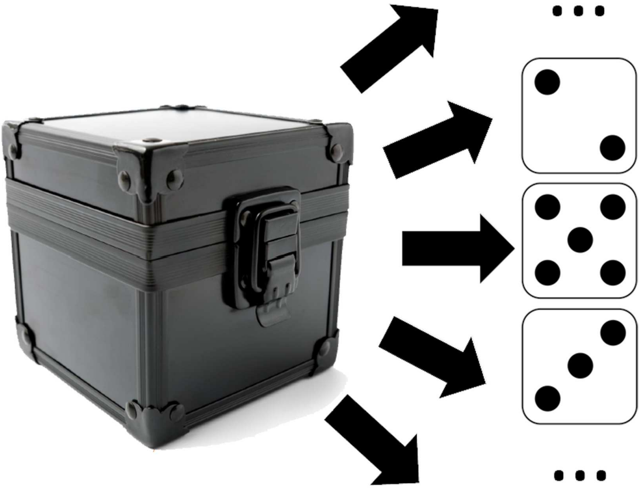
Случайность



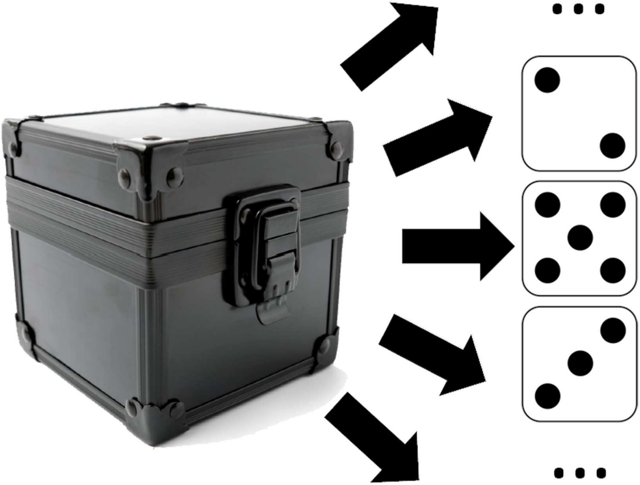
Случайность

Случайность

Изучение случайности

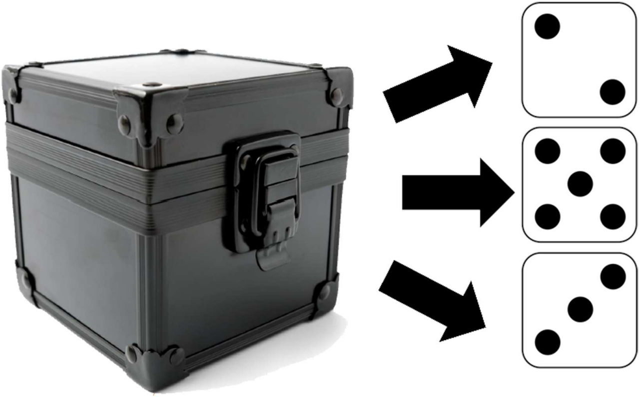


Изучение случайности



Вероятность события — доля испытаний, завершившихся наступлением события, в бесконечном эксперименте.

Изучение случайности



Изучение случайности

Изучение случайности

Изучение случайности

Закон больших чисел: на больших выборках частота события хорошо приближает его вероятность.

Дискретные случайные величины

Пусть величина X принимает значения, которые можно перенумеровать. Поставим каждому в соответствие вероятность его появления:

$$\{x_1, x_2, x_3, \dots\} \quad \Longrightarrow \quad \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ \dots \end{pmatrix},$$

причем:

$$p_i \geq 0, \qquad \sum_{i=1} p_i = 1.$$

$\mathbf{P}(X = x_i) = p_i$ — функция вероятности.

Распределение Бернулли

Случайная величина X с двумя исходами:

$$\mathbf{P}(X = 1) = p,$$
$$\mathbf{P}(X = 0) = 1 - p.$$

Биномиальное распределение

X — сумма n одинаковых бернуллиевских случайных величин с параметром p:

$$\mathbf{P}(X=0)=(1-p)^n\,,$$
$$\mathbf{P}(X=n)=p^n,$$
$$\mathbf{P}(X=k)=\frac{n!}{k!\,(n-k)!}p^k\,(1-p)^{n-k}\,,k=0,1,\ldots,n.$$

Распределение Пуассона

X — счётчик:

$$\mathbf{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \lambda > 0, k = 0, 1, \ldots$$

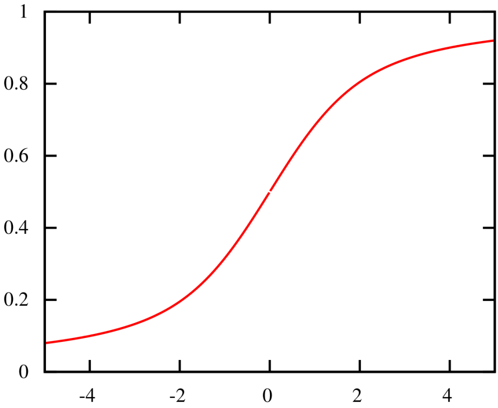
то есть спустя три года, скромницей, с чемоданчиком балерины в руке, затем — шестнадцати лет, в пачках, с газовыми крыльцами за спиной, вольно сидящей на столе, с поднятым бокалом, среди бледных гуляк, затем — лет восемнадцати, в фатальном трауре, у перил над каскадом, затем... ах, во многих еще видах и позах, вплоть до самой последней — лежачей.

При помощи ретушировки и других фотофокусов как будто достигалось последовательное изменение лица Эммочки (искусник, между прочим, пользовался фотографиями ее матери), но стоило взглянуть ближе, и становилась безобразно ясной аляповатость этой пародии на работу времени. У Эммочки, выходившей из театра в мехах с цветами, прижатыми к плечу, были ноги, никогда не плясавшие; а на следующем снимке, изображавшем ее уже в венчальной дымке, стоял рядом с ней жених, стройный и высокий, но с кругленькой физиономией м-сье Пьера. В тридцать лет у нее появились условные морщины, проведенные без смысла, без жизни, без знания их истинного значения, — но знатоку говорящие совсем странное, как бывает, что случайное движение ветвей совпадает с жестом, понятным для глухонемого. А в сорок лет

Непрерывные случайные величины

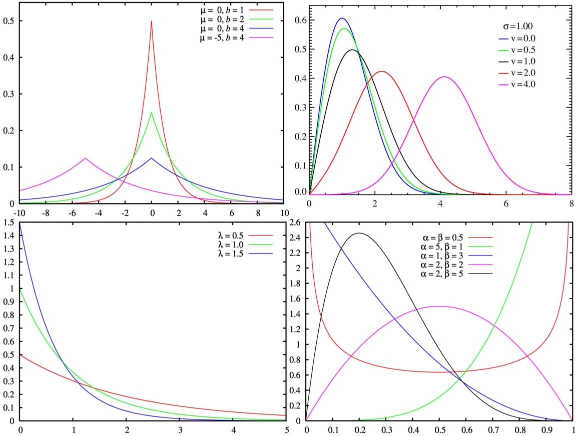
Если множество значений X нельзя перенумеровать (например, $[0,1]$), то её распределение нельзя задать с помощью функции вероятности, потому что $\mathbf{P}(X = x) = 0 \;\; \forall x$.

$F_X(x) = \mathbf{P}(X \leqslant x)$ — функция распределения.



Непрерывные случайные величины

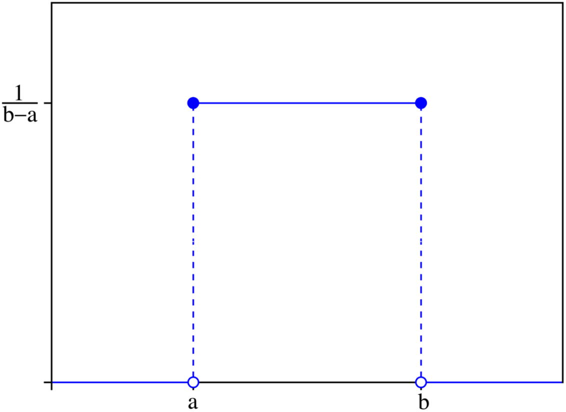
$f\left(x\right):\ \int_a^b f\left(x\right)dx = \mathbf{P}(a \leqslant X \leqslant b)$ — плотность распределения.
$F\left(x\right) = \int_{-\infty}^{x} f\left(u\right)du$



Непрерывное равномерное распределение

$X \sim U\,[a,b]:$

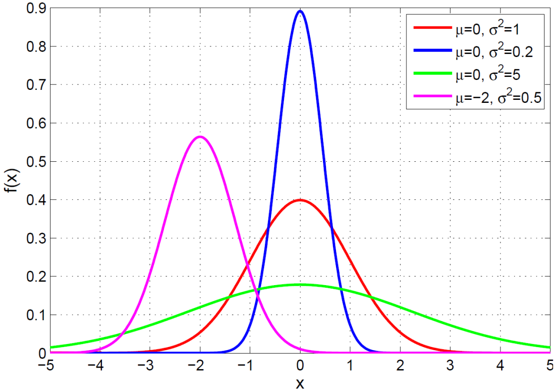
$$f\,(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b]\,, \\ 0, & x \notin [a,b]\,. \end{cases}$$



Нормальное распределение

$X \sim N\left(\mu, \sigma^2\right)$:

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



Выборка

Генеральная совокупность — множество объектов, свойства которых подлежат изучению в рассматриваемой задаче.

Выборка — конечное множество объектов, отобранных из генеральной совокупности для проведения измерений.

$$X^n = (X_1, \ldots, X_n) .$$

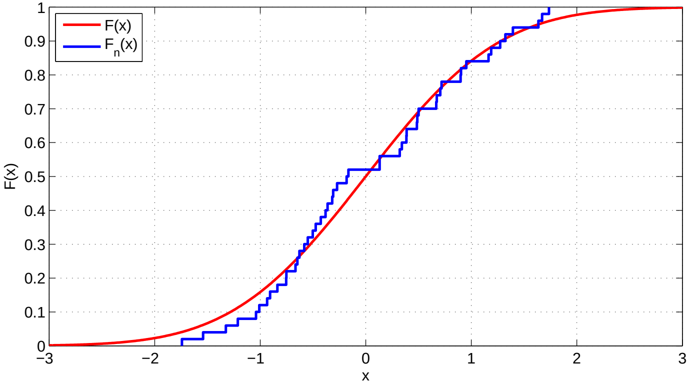
n — **объём выборки**.

X^n — **простая выборка**, если $X_1, \ldots, X_n$ — независимые одинаково распределённые случайные величины (i.i.d.).

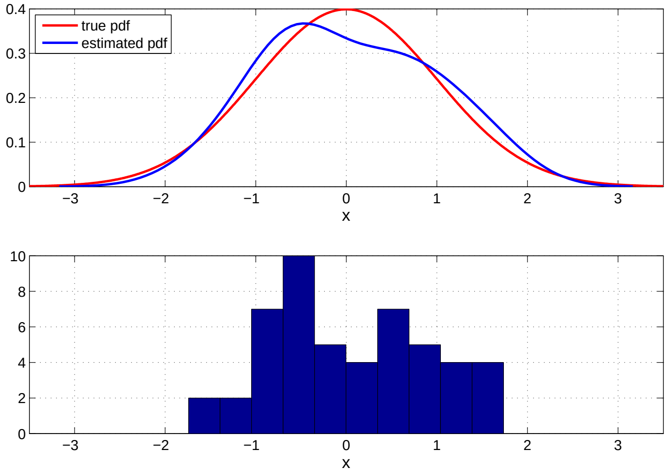
Основная задача статистики — описание $F_X(x)$ по реализации выборки.

Функция распределения

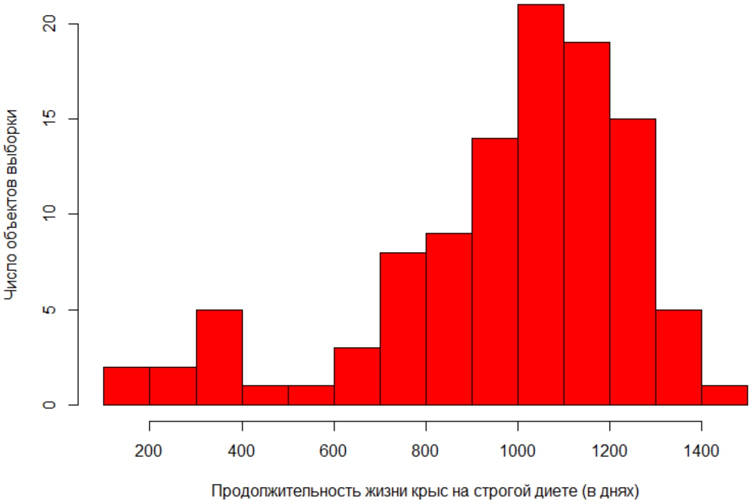
$F_n(x) = \frac{1}{n}\sum\limits_{i=1}^{n}[X_i \leqslant x]$ — **эмпирическая функция распределения**.



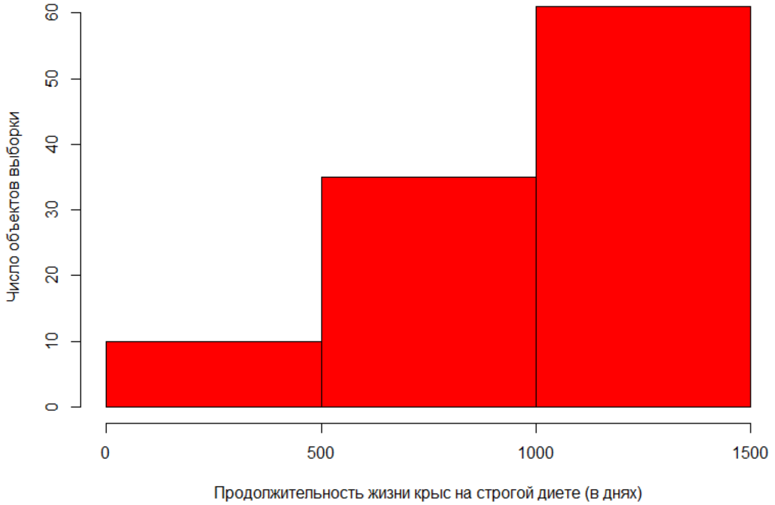
Плотность распределения



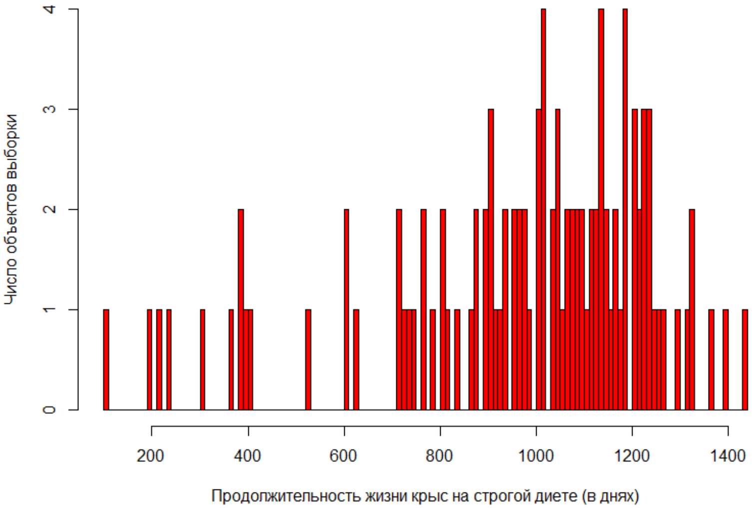
Гистограммы



Гистограммы



Гистограммы



Характеристики распределений

- **матожидание** — среднее значение X:

$$\mathbb{E}X = \int x\,dF(x);$$

- **дисперсия** — мера разброса X:

$$\mathbb{D}X = \mathbb{E}\left((X - \mathbb{E}X)^2\right);$$

- **квантиль** порядка $\alpha \in (0,1)$:

$$X_\alpha\colon \quad \mathbf{P}(X \leqslant X_\alpha) \geqslant \alpha, \;\; \mathbf{P}(X \geqslant X_\alpha) \geqslant 1 - \alpha;$$

  эквивалентное определение:

$$X_\alpha = F^{-1}(\alpha) = \inf\{x\colon F(x) \geqslant \alpha\}.$$

- **медиана** — квантиль порядка 0.5, центральное значение распределения;
- **мода** — точка максимума функции вероятности или плотности:

$$\mathrm{mode}\, X = \operatorname*{argmax}_x f(x);$$

Статистика

Статистика $T(X^n)$ — любая измеримая функция выборки.

- выборочное среднее:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

- выборочная дисперсия:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2;$$

вариационный ряд:

$$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)};$$

ранг элемента выборки X_i:

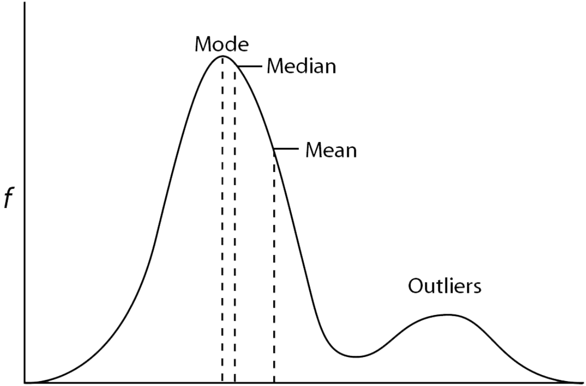
$$\operatorname{rank}\left(X_i\right) = r\colon X_i = X_{(r)};$$

- k-я порядковая статистика: $X_{(k)}$;
- выборочный α-квантиль: $X_{([n\alpha])}$;
- выборочная медиана:

$$m = \begin{cases} X_{(k+1)}, & \text{если } n = 2k+1, \\ \frac{X_{(k)}+X_{(k+1)}}{2}, & \text{если } n = 2k. \end{cases}$$

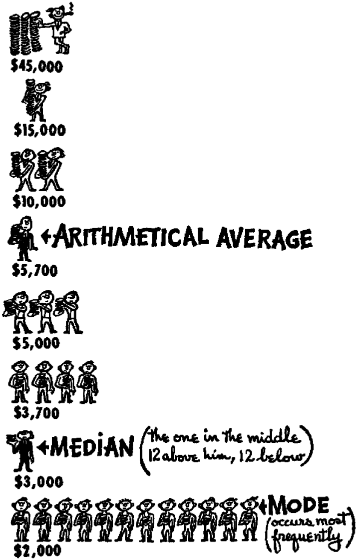
Оценки центральной тенденции

Выборочное среднее — среднее арифметическое по выборке.
Выборочная медиана — центральный элемент вариационного ряда.
Выборочная мода — самое распространённое значение в выборке.

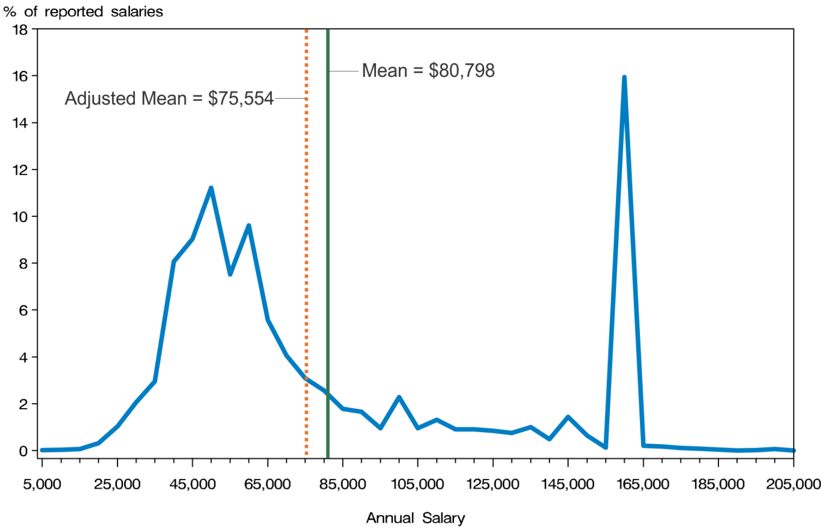


Оценки центральной тенденции

(Huff, 1954):



Об ограниченности статистик



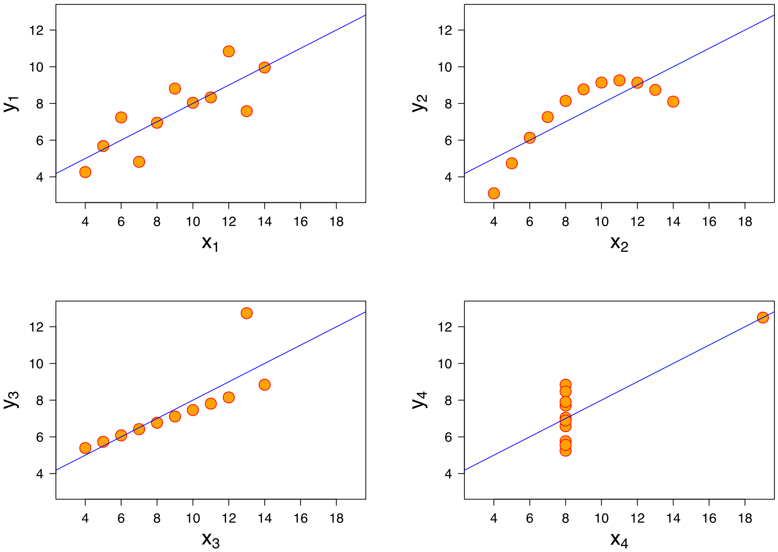
Уровень стартовой заработной платы выпускников юридических факультетов, США, 2012, данные NALP.

Об ограниченности статистик

Квартет Энскомба (Anscombe, 1973):

№	1	2	3	4
$\bar{x}$	9	9	9	9
S_x	11	11	11	11
$\bar{y}$	7.5	7.5	7.5	7.5
S_y	4.127	4.127	4.128	4.128
r_{xy}	0.816	0.816	0.816	0.816

Об ограниченности статистик



Точечные оценки

Пусть распределение генеральной совокупности параметрическое:

$$F(x) = F(x, \theta).$$

Статистика $\hat{\theta}_n = \hat{\theta}(X^n)$ — точечная оценка параметра θ.
Какая оценка лучше?

Состоятельность: $\lim\limits_{n\to\infty} \mathbf{P}\left(\hat{\theta}_n = \theta\right) = 1.$

Несмещённость: $\mathbb{E}\hat{\theta}_n = \theta.$

Асимптотическая несмещённость: $\lim\limits_{n\to\infty} \mathbb{E}\hat{\theta}_n = \theta.$

Оптимальность: $\mathbb{D}\hat{\theta}_n = \min\limits_{\hat{\theta}\,:\ \mathbb{E}\hat{\theta}=\theta} \mathbb{D}\hat{\theta}.$

Робастность: устойчивость $\hat{\theta}_n$ относительно

- отклонений истинного распределения X от модельного семейства;
- выбросов, содержащихся в выборке.

Интервальные оценки

Доверительный интервал:

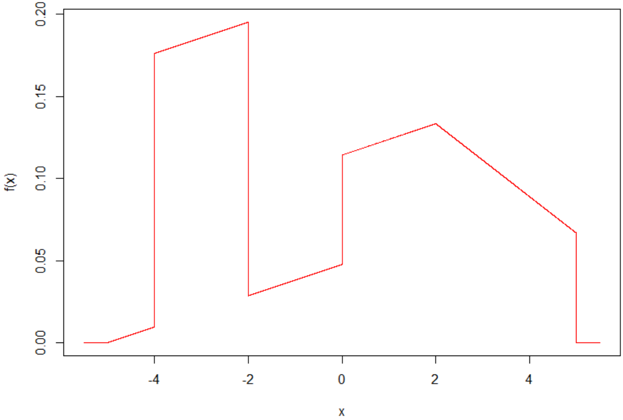
$$\mathbf{P}(\theta \in [C_L, C_U]) \geqslant 1 - \alpha,$$

$1 - \alpha$ — уровень доверия,
C_L, C_U — нижний и верхний доверительные пределы.

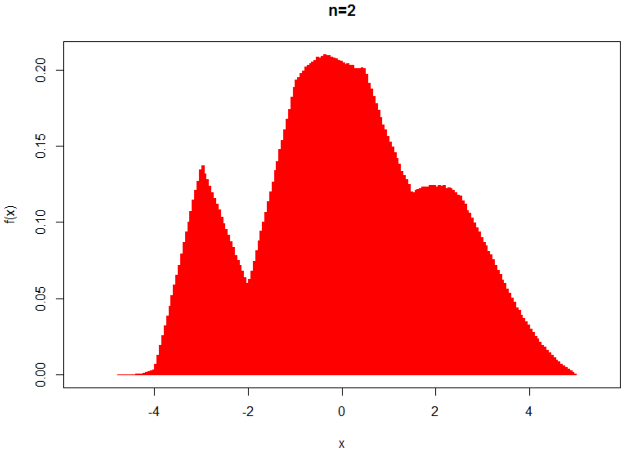
Неверная интерпретация: неизвестный параметр лежит в пределах построенного доверительного интервала с вероятностью $1 - \alpha$.

Верная интерпретация: при бесконечном повторении процедуры построения доверительного интервала на аналогичных выборках в $100(1 - \alpha)\%$ случаев он будет содержать истинное значение θ.

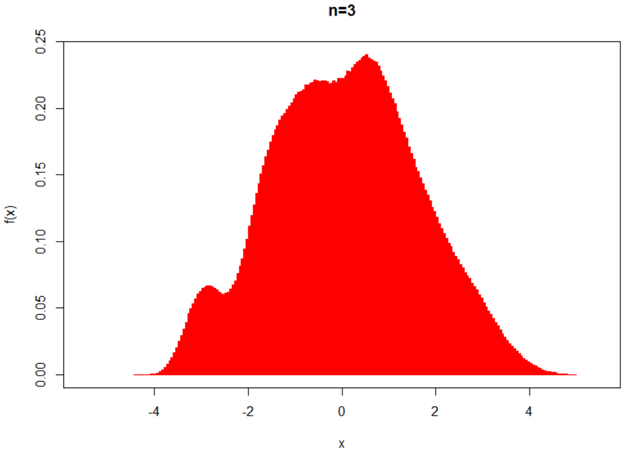
Странное распределение



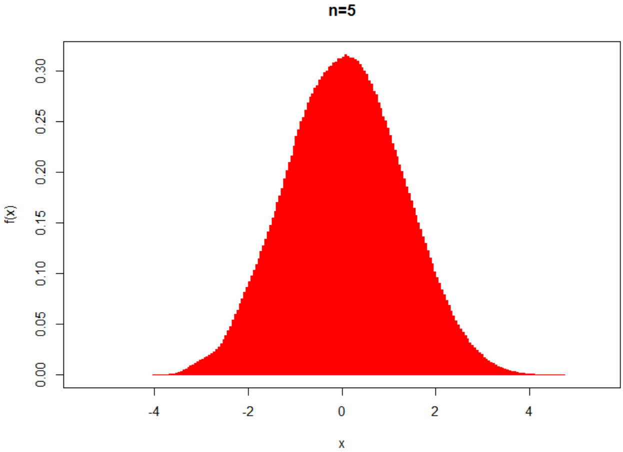
Странное распределение



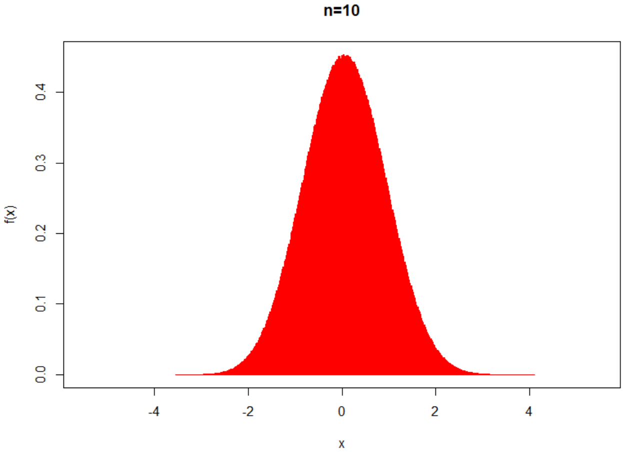
Странное распределение



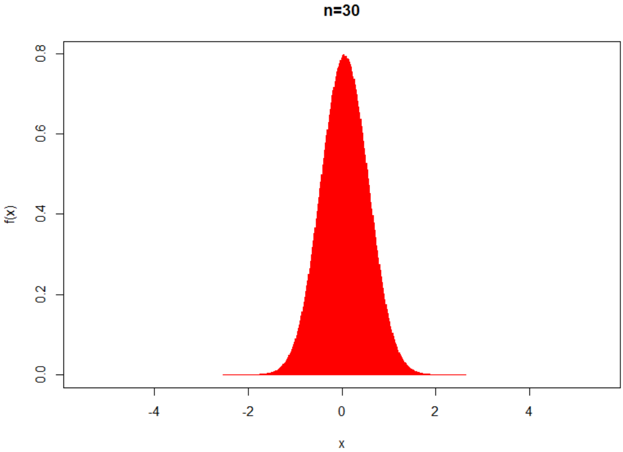
Странное распределение



Странное распределение



Странное распределение

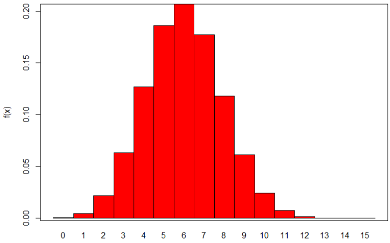


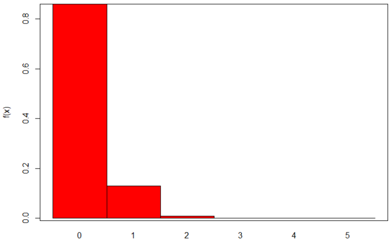
Центральная предельная теорема

Пусть $X_1, \dots, X_n$ i.i.d. с $\mathbb{E}X$ и $\mathbb{D}X < \infty$, тогда

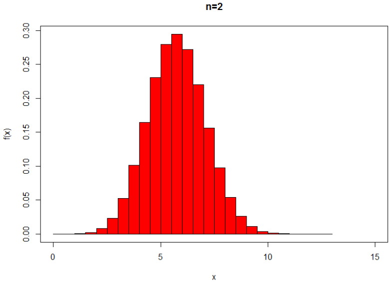
$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim\approx N\left(\mathbb{E}X, \frac{\mathbb{D}X}{n}\right).$$

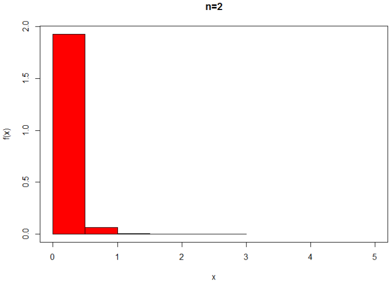
Точность аппроксимации



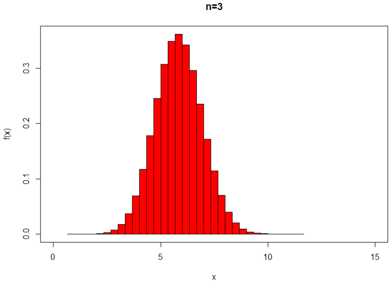


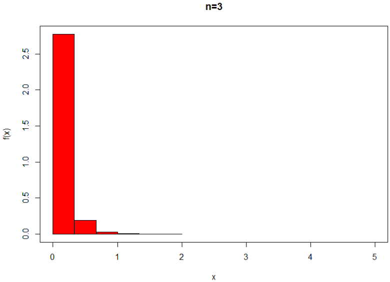
Точность аппроксимации



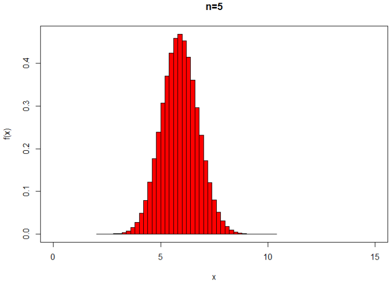


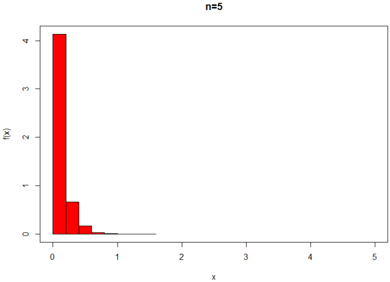
Точность аппроксимации



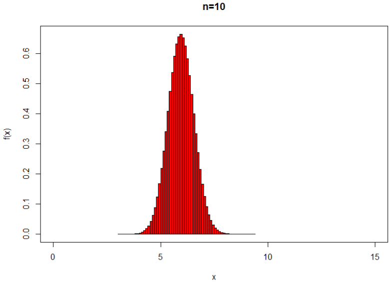


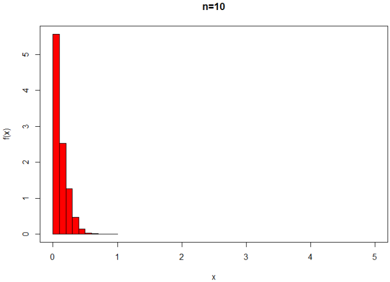
Точность аппроксимации



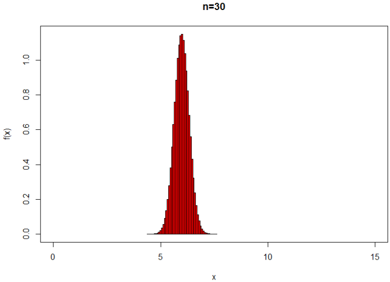


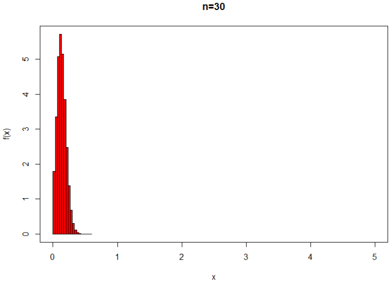
Точность аппроксимации





Точность аппроксимации





Доверительный интервал

$X \sim F(x, \theta)$, θ — неизвестный параметр,
$\theta = ?$

$X^n = (X_1, \ldots, X_n)$,
$\hat{\theta}$ — оценка θ по выборке.
Если мы знаем распределение $\hat{\theta}$ $F_{\hat{\theta}}(x)$, то:

$$\mathbf{P}\left(F_{\hat{\theta}}^{-1}\left(\frac{\alpha}{2}\right) \leqslant \theta \leqslant F_{\hat{\theta}}^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha.$$

Доверительный интервал для среднего

$X \sim F(x), \quad X^n = (X_1, \dots, X_n)$,
$\bar{X}_n$ — оценка $\mathbb{E}X$,

$\bar{X}_n \approx\sim N\left(\mathbb{E}X, \frac{\mathbb{D}X}{n}\right)$ (ЦПТ) $\Rightarrow$

доверительный интервал для $\mathbb{E}X$:

$$\mathbf{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathbb{D}X}{n}} \leqslant \mathbb{E}X \leqslant \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathbb{D}X}{n}}\right) \approx 1-\alpha.$$

Задача

Имеется продукт, определена его целевая аудитория. Как оценить его узнаваемость?

$$X = \begin{cases} 1, & \text{член ЦА знает продукт,} \\ 0, & \text{не знает.} \end{cases}$$

$\hat{p}_n = \bar{X}_n$

Опрос 1: $n = 10, \hat{p}_n = 0.6$
Опрос 2: $n = 100, \hat{p}_n = 0.44$
ЦПТ: $\hat{p}_n = \bar{X}_n \sim\approx N\left(p, \frac{p(1-p)}{n}\right) \approx N\left(\hat{p}_n, \frac{\hat{p}_n(1-\hat{p}_n)}{n}\right)$
Правило двух сигм:

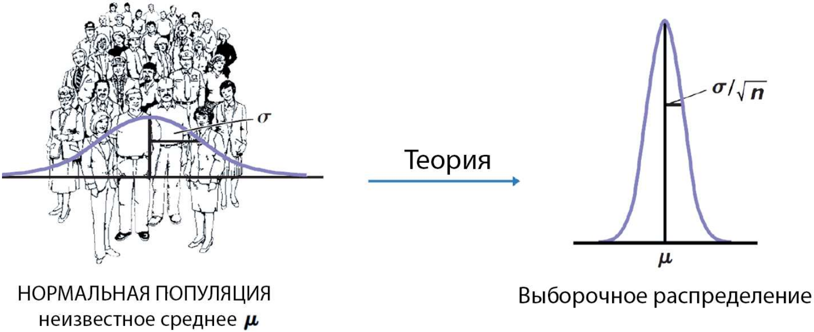
$$\mathbf{P}\left(\hat{p}_n - 2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \leqslant p \leqslant \hat{p}_n + 2\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right) \approx 0.95 \Rightarrow$$

Опрос 1: $p \in [0.29; 0.91]$
Опрос 2: $p \in [0.34; 0.54]$

Построение доверительных интервалов

Как можно оценить $F_{\hat{\theta}_n}(x)$ — выборочное распределение статистики $\hat{\theta}_n$? (Hesterberg, 2005):

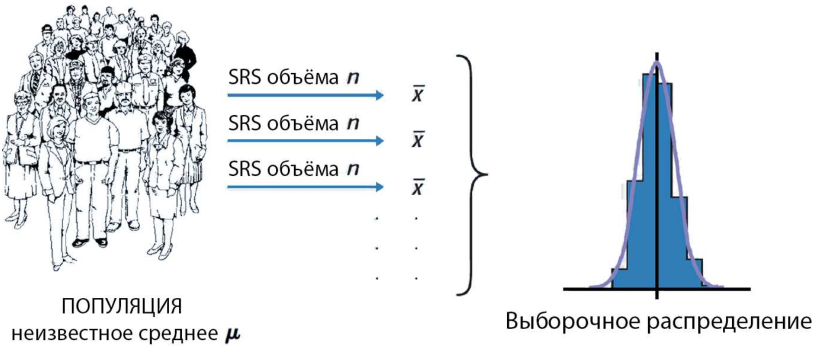
- параметрический метод:



Сделать предположение, что X распределена по закону $F_X(x)$, при выполнении которого закон распределения $\hat{\theta}_n$ известен.

Построение доверительных интервалов

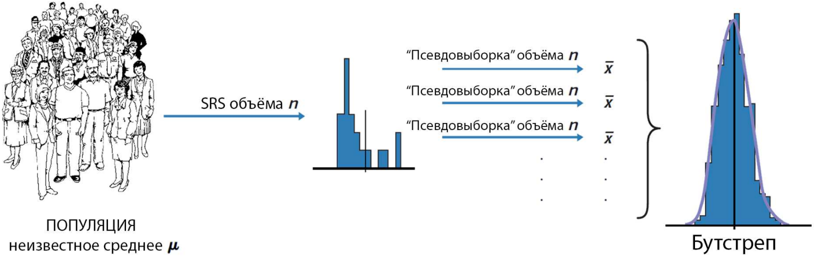
- наивный метод:



Извлечь из генеральной совокупности N выборок объёма n и оценить выборочное распределение $\hat{\theta}_n$ эмпирическим.

Построение доверительных интервалов

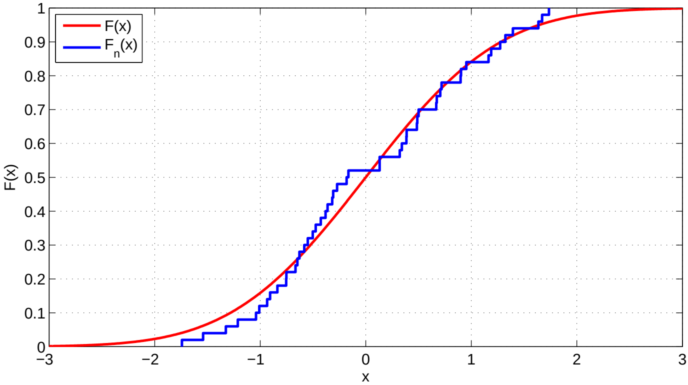
- бутстреп:



Сгенерировать N «псевдовыборок» объёма n и оценить выборочное распределение $\hat{\theta}_n$ «псевдоэмпирическим».

Бутстреп

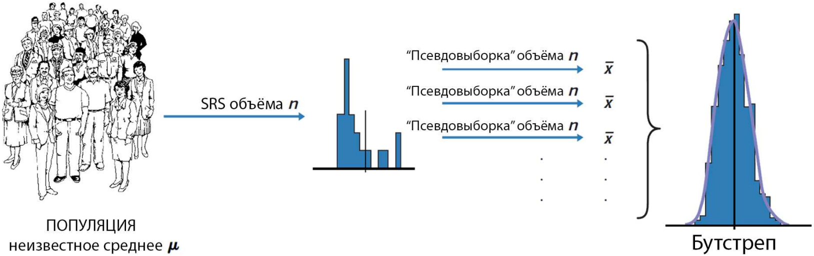
Извлечение выборок из генеральной совокупности — сэмплирование из неизвестного распределения $F_X(x)$.
Лучшая оценка $F_X(x)$, которая у нас есть — $F_{X^n}(x)$:



Сэмплировать из неё — это то же самое, что делать из X^n выборки с возвращением объёма n.

Бутстреп-распределение

$X^{1*}, \dots, X^{N*}$ — бутстреп-псевдовыборки из X^n объёма n,
$\hat{\theta}_n^{1*}, \dots, \hat{\theta}_n^{N*}$ — значения статистики на них,
$F_{\hat{\theta}_n}^{boot}(x)$ — бутстреп-распределение $\hat{\theta}_n$ — эмпирическая функция распределения, построенная по значениям статистики на псевдовыборках.



По $F_{\hat{\theta}_n}^{boot}(x)$ можно строить доверительные интервалы для θ!

Доверительные интервалы

- Посчитаем S_n^{boot} — выборочное стандартное отклонение $\hat{\theta}_n$ на псевдовыборках;

$$\mathbf{P}\left(\hat{\theta}_n - t_{n-1,1-\frac{\alpha}{2}} S_n^{boot} \leqslant \theta \leqslant \hat{\theta}_n + t_{n-1,1-\frac{\alpha}{2}} S_n^{boot}\right) \approx 1-\alpha.$$

  Это стьюдентизированный бутстреп.

- Возьмём выборочные квантили бутстреп-распределения:

$$\mathbf{P}\left(\left(F_{\hat{\theta}_n}^{boot}\right)^{-1}\left(\frac{\alpha}{2}\right) \leqslant \theta \leqslant \left(F_{\hat{\theta}_n}^{boot}\right)^{-1}\left(1-\frac{\alpha}{2}\right)\right) \approx 1-\alpha.$$

  Это базовый бутстреп.

Литература

Справочники по статистике:

- Кобзарь А.И. *Прикладная математическая статистика*, 2006.
- Kanji G.K. *100 statistical tests*, 2006.

Вводные учебники по статистике:

- Good P.I., Hardin J.W. *Common Errors in Statistics (and How to Avoid Them)*, 2003.
- Reinhart A. *Statistics Done Wrong. The woefully complete guide*, `http://www.statisticsdonewrong.com/`

Бутстреп:

- Hesterberg T., Monaghan S., Moore D.S., Clipson A., Epstein R. *Bootstrap methods and permutation tests*. In Introduction to the Practice of Statistics, 2005. `http://statweb.stanford.edu/~tibs/stat315a/Supplements/bootstrap.pdf`
- Efron B., Tibshirani R. *An Introduction to the Bootstrap*, 1993.