

# Прикладной статистический анализ данных. 1. Введение.

Рябенко Евгений  
riabenko.e@gmail.com

I/2015

# Выборка

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

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

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

$n$  — **объём выборки**.

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

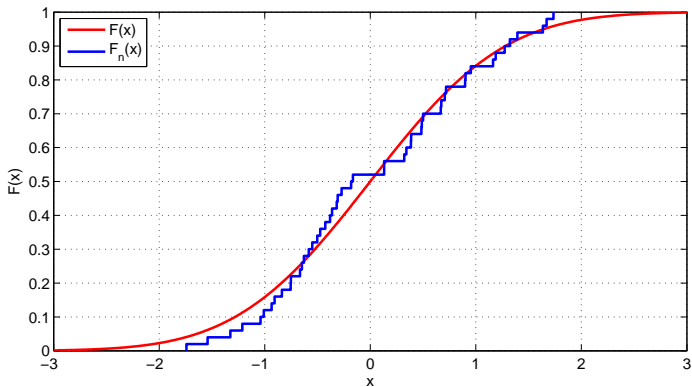
Пусть  $F(x)$  — функция распределения элемента простой выборки:

$$F(x) = \mathbf{P}(X \leq x).$$

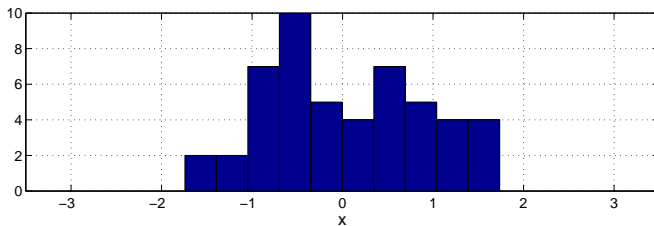
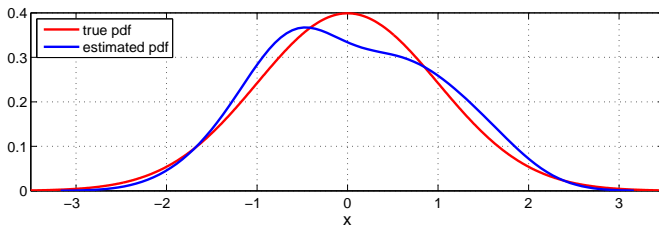
Основная задача статистики — описание  $F(x)$  по реализации выборки.

## Функция распределения

$F_n(x) = \frac{1}{n} \sum_{i=1}^n [X_i \leq x]$  — эмпирическая функция распределения.



# Плотность распределения



## Характеристики распределения

Часто интерес представляют отдельные характеристики распределения  $F(x)$ :

- **матожидание** — среднее значение  $X$ :

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

- **дисперсия** — мера разброса  $X$ :

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

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

$$X_\alpha: \quad \mathbf{P}(X \leq X_\alpha) \geq \alpha, \quad \mathbf{P}(X \geq X_\alpha) \geq 1 - \alpha;$$

- **медиана** — квантиль порядка 0.5, центральное значение распределения:

$$\text{med } X: \quad \mathbf{P}(X \leq \text{med } X) \leq \frac{1}{2}, \quad \mathbf{P}(X \geq \text{med } X) \geq \frac{1}{2};$$

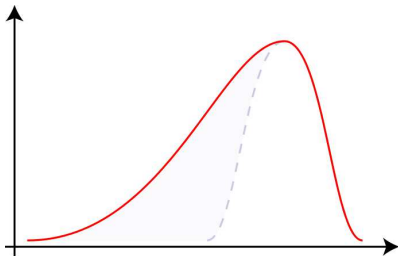
- **интерквартильный размах**:

$$IQR = X_{3/4} - X_{1/4};$$

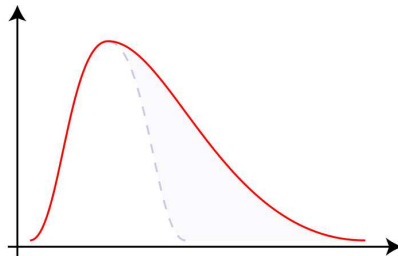
# Характеристики распределения

- коэффициент асимметрии (skewness):

$$\gamma_1 = \mathbb{E} \left( \frac{X - \mathbb{E}X}{\sqrt{\mathbb{D}X}} \right)^3 ;$$



Negative Skew

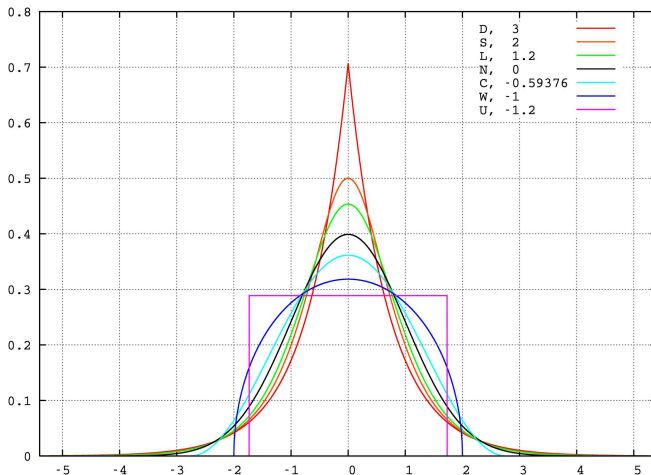


Positive Skew

# Характеристики распределения

- коэффициент эксцесса (excess, без вычитания трёх — kurtosis):

$$\gamma_2 = \frac{\mathbb{E}(X - \mathbb{E}X)^4}{(\mathbb{D}X)^2} - 3.$$



## Статистика

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

- выборочное среднее:

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

- выборочная дисперсия:

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

**вариационный ряд:**

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)};$$

**ранг** элемента выборки  $X_i$ :

$$\text{rank}(X_i) = \mathbb{E} \{ r \mid X_i = X_{(r)} \};$$

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

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



- выборочный интерквартильный размах:

$$IQR_n = X_{([3n/4])} - X_{([n/4])};$$

- выборочный коэффициент асимметрии:

$$g_1 = \frac{\sqrt{n} \sum_{i=1}^n (X_i - \bar{X})^3}{\left( \sum_{i=1}^n (X_i - \bar{X})^2 \right)^{3/2}};$$

- выборочный коэффициент эксцесса:

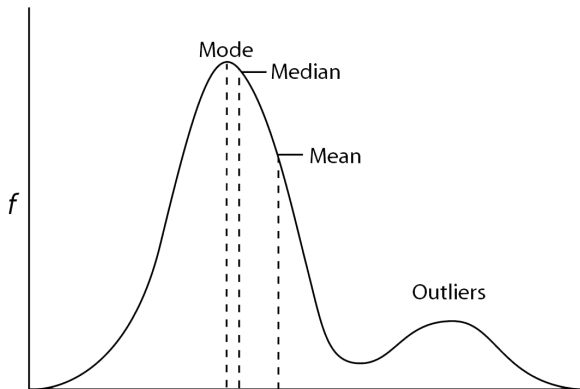
$$g_2 = \frac{n \sum_{i=1}^n (X_i - \bar{X})^4}{\left( \sum_{i=1}^n (X_i - \bar{X})^2 \right)^2} - 3.$$

## Оценки центральной тенденции

Выборочное среднее — среднее арифметическое по выборке.

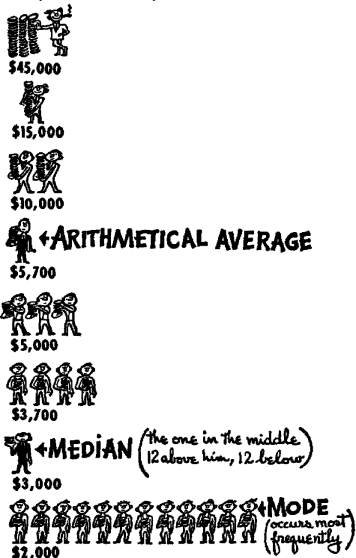
Медиана — центральный элемент вариационного ряда.

Мода — самое распространённое значение в выборке.

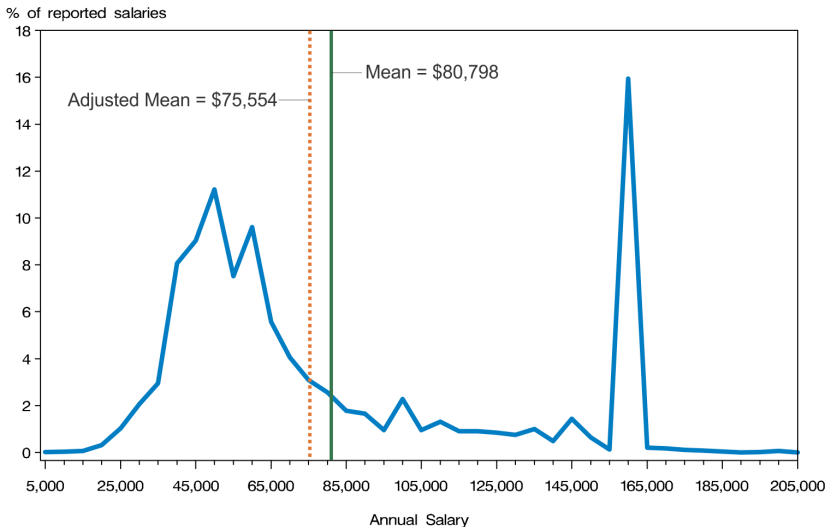


# Оценки центральной тенденции

How to lie with statistics (Huff, 1954)



# Об ограниченности статистик



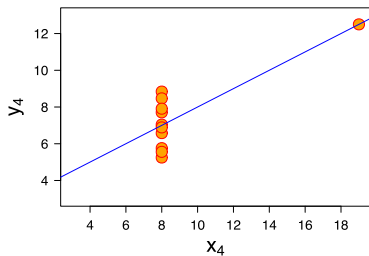
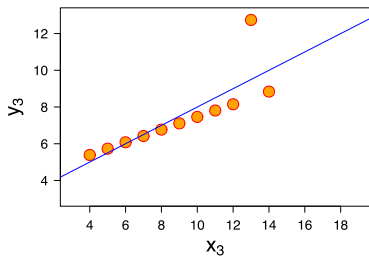
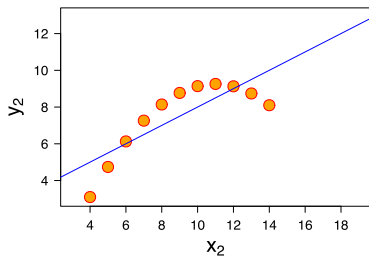
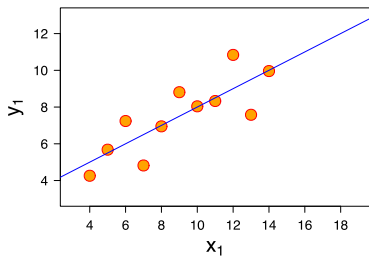
Уровень стартовой заработной платы выпускников юридических факультетов, США, 2012, данные NALP.

## Об ограниченности статистик

Квартет Энскомба (Anscombe's quartet):

№	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:

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

$\hat{\theta}_n = \hat{\theta}(X^n)$  — статистика, точечная оценка параметра.

Какая оценка лучше?

**Состоятельность:**  $\text{plim}_{n \rightarrow \infty} \hat{\theta}_n = \theta.$

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

**Асимптотическая несмещённость:**  $\lim_{n \rightarrow \infty} \mathbb{E}\hat{\theta}_n = \theta.$

**Оптимальность:**  $\mathbb{D}\hat{\theta}_n = \min_{\hat{\theta}: \mathbb{E}\hat{\theta}=\theta} \mathbb{D}\hat{\theta}.$

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

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

## Интервальные оценки

Оценим параметр  $\theta$  двумя статистиками:

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

$1 - \alpha$  — уровень доверия,

$C_L, C_U$  — верхний и нижний 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 параметра.



## Интервальные оценки

**Пример 1:** доверительный интервал для среднего  $X \sim N(\mu, \sigma^2)$  при известной дисперсии  $\sigma^2$ .

$$X^n = (X_1, \dots, X_n), \quad X_i \sim N(\mu, \sigma^2) \Rightarrow$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim N(0, 1) \Rightarrow$$

$$\mathbf{P}\left(\mu \in \left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha.$$

$z_{1-\frac{\alpha}{2}}$  —  $(1 - \frac{\alpha}{2})$ -квантиль стандартного нормального распределения; при  $\alpha = 0.05$  получаем  $z_{0.975} \approx 1.96$ .

Правило двух сигм: если  $X \sim N(\mu, \sigma^2)$ , то  $\mathbf{P}(|X - \mu| \leq 2\sigma) \approx 0.954$ .

Если  $X$  распределена не нормально, то можно утверждать только  $\mathbf{P}(|X - \mathbb{E}X| \leq 2\sqrt{\mathbb{D}X}) \geq 0.75$  (из неравенства Чебышёва).

## Интервальные оценки

**Пример 2:** непараметрический доверительный интервал для медианы непрерывного распределения.

$$X^n = (X_1, \dots, X_n), \quad X_i \sim F(x) \Rightarrow$$

$$\mathbf{P}(\text{med } X_i \in [X_{(r)}, X_{(n-r+1)}]) = \frac{1}{2^n} \sum_{i=r}^{n-r+1} C_n^i.$$

При  $n > 10$  применима нормальная аппроксимация:

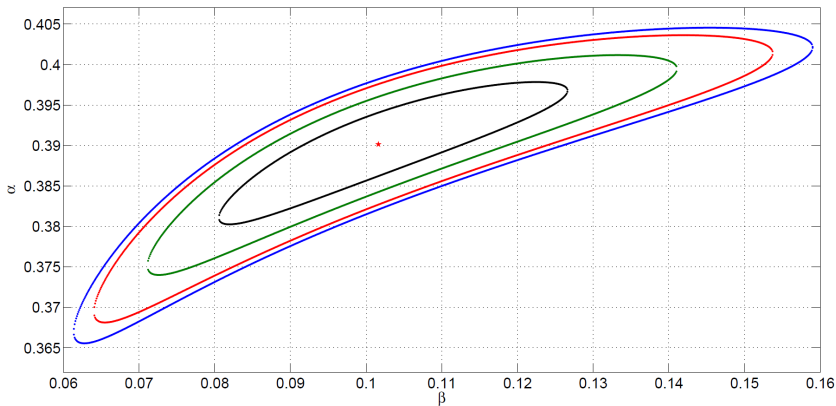
$$\mathbf{P}\left(\text{med } X_i \in \left[ X\left(\left[\frac{n-\sqrt{n}z_{1-\frac{\alpha}{2}}}{2}\right]\right), X\left(\left[\frac{n+\sqrt{n}z_{1-\frac{\alpha}{2}}}{2}\right]\right) \right]\right) \approx 1 - \alpha.$$

Аналогично строится непараметрический доверительный интервал для любого квантиля  $X_p, p \in (0, 1)$ :

$$\mathbf{P}(X_p \in [X_{(l)}, X_{(u)}]) = \sum_{i=l}^u C_n^i p^i (1-p)^{n-i}.$$

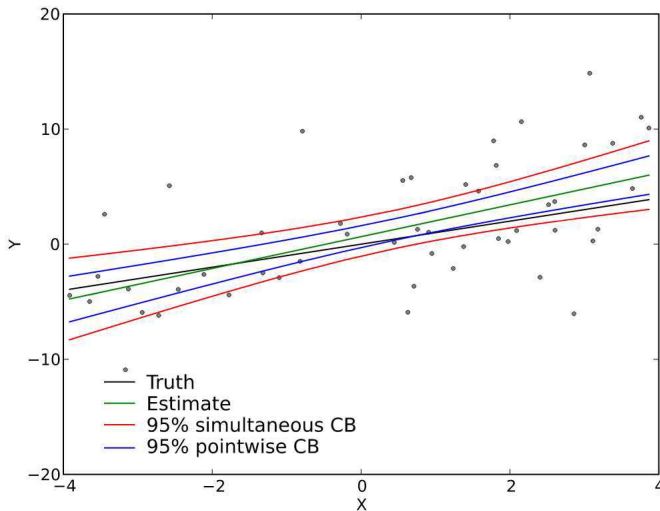
# Интервальные оценки

Доверительная область для пары неизвестных параметров  $(\alpha, \beta)$ :



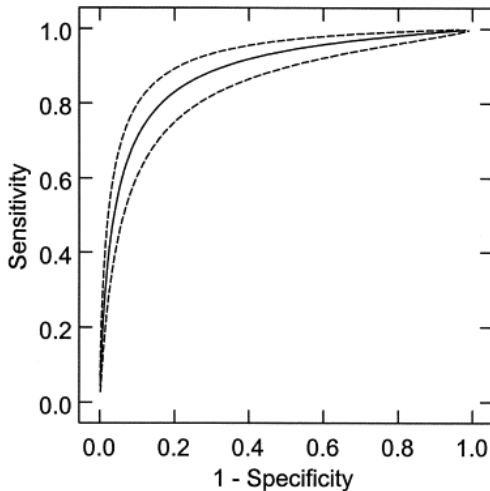
# Интервальные оценки

Доверительная лента для функции  $Y = \beta_0 + \beta_1 x$ :



# Интервальные оценки

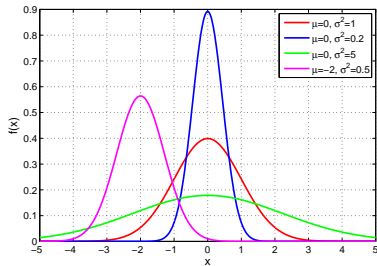
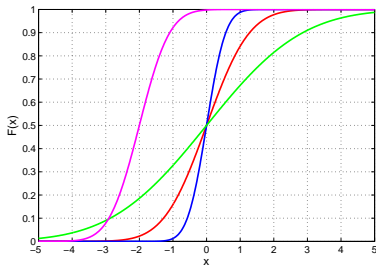
Доверительная лента для ROC-кривой:



# Нормальное распределение

$X \in \mathbb{R} \sim N(\mu, \sigma^2)$ ,  $\sigma^2 > 0$  — предельное распределение суммы слабо  
взаимозависимых сл. в.

Пример: погрешность измерения.



$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right),$$

$$f(x) = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right),$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt,$$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

# Нормальное распределение

$$\mathbb{E}X = \mu,$$

$$\text{med } X = \mu,$$

$$\text{mode } X = \mu,$$

$$\mathbb{D}X = \sigma^2,$$

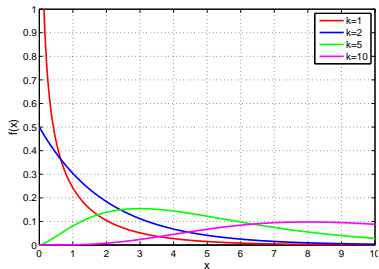
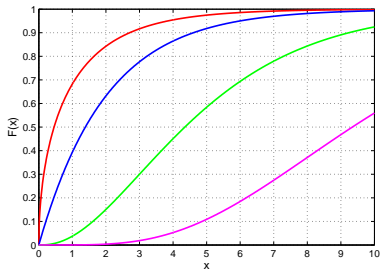
$$\gamma_1(X) = 0,$$

$$\gamma_2(X) = 0.$$

# Распределение хи-квадрат

$X \in \mathbb{R}_+ \sim \chi_k^2$ ,  $k \in \mathbb{N}$  — распределение суммы квадратов  $k$  независимых стандартных нормальных сл. в.

Пример: выборочная дисперсия.



$$F(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right),$$

$$f(x) = \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}.$$

$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$  — гамма-функция,

$\gamma(a, x) = \int_0^x e^{-t} t^{a-1} dt$  — нижняя неполная гамма-функция.



## Распределение хи-квадрат

$$\mathbb{E}X = k,$$

$$\text{med } X \approx k \left(1 - \frac{2}{9k}\right)^3,$$

$$\text{mode } X = \max(k - 2, 0),$$

$$\mathbb{D}X = 2k,$$

$$\gamma_1(X) = \sqrt{8/k},$$

$$\gamma_2(X) = 12/k.$$

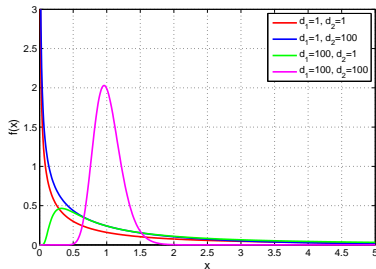
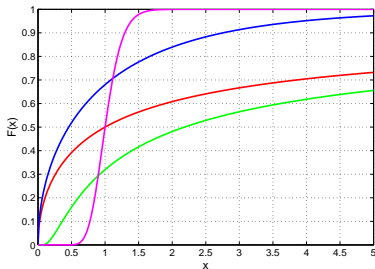
- Пусть  $X_1, \dots, X_k$  — i.i.d.,  $X_i \sim N(0, 1)$ , тогда

$$\sum_{i=1}^k X_i^2 \sim \chi_k^2.$$

# Распределение Фишера

$X \in \mathbb{R}_+ \sim F(d_1, d_2)$ ,  $d_1, d_2 > 0$  — распределение отношения двух независимых нормированных хи-квадрат сл. в.

Возникает в дисперсионном и регрессионном анализе.



$$F(x) = I_{\frac{d_1 x}{d_1 x + d_2}} \left( \frac{d_1}{2}, \frac{d_2}{2} \right),$$

$$f(x) = \sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}} / x B \left( \frac{d_1}{2}, \frac{d_2}{2} \right).$$

$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$  — бета-функция,

$I_x(a, b) = \frac{B(x; a, b)}{B(a, b)}$  — регуляризованная неполная бета-функция,

$B(x; a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt$  — неполная бета-функция.

## Распределение Фишера

$$\mathbb{E}X = \frac{d_2}{d_2 - 2} \text{ при } d_2 > 2,$$

$$\text{mode } X = \frac{d_1 - 2}{d_1} \frac{d_2}{d_2 + 2} \text{ при } d_1 > 2,$$

$$\mathbb{D}X = \frac{2d_2^2 (d_1 + d_2 - 2)}{d_1 (d_2 - 2)^2 (d_2 - 4)} \text{ при } d_2 > 4,$$

$$\gamma_1(X) = \frac{(2d_1 + d_2 - 2) \sqrt{8(d_2 - 4)}}{(d_2 - 6) \sqrt{d_1 (d_1 + d_2 - 2)}} \text{ при } d_2 > 6,$$

$$\gamma_2(X) = 12 \frac{d_1(5d_2 - 22)(d_1 + d_2 - 2) + (d_2 - 4)(d_2 - 2)^2}{d_1(d_2 - 6)(d_2 - 8)(d_1 + d_2 - 2)} \text{ при } d_2 > 8.$$

## Распределение Фишера

- Пусть  $X_1 \sim \chi_{d_1}^2$ ,  $X_2 \sim \chi_{d_2}^2$ ,  $X_1$  и  $X_2$  независимы, тогда

$$\frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2).$$

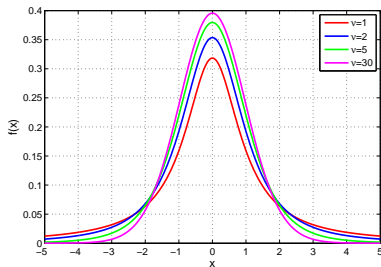
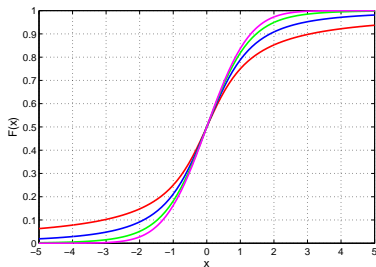
- Если  $X \sim F(d_1, d_2)$ , то

$$Y = \lim_{d_2 \rightarrow \infty} d_1 X \sim \chi_{d_1}^2.$$

- $F(x, d_1, d_2) = F(1/x, d_2, d_1)$ .

# Распределение Стьюдента

$X \in \mathbb{R} \sim St(\nu)$ ,  $\nu > 0$  — распределение отношения независимых стандартной нормальной сл. в. и корня из нормированной хи-квадрат сл. в. Возникает при оценке среднего значения сл. в. с неизвестной дисперсией.



$$F(x) = \frac{1}{2} + x \Gamma\left(\frac{\nu+1}{2}\right),$$

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

## Распределение Стьюдента

$$\mathbb{E}X = 0 \text{ при } \nu > 1,$$

$$\text{med } X = 0,$$

$$\text{mode } X = 0,$$

$$\mathbb{D}X = \begin{cases} \frac{\nu}{\nu-2}, & \nu > 2, \\ \infty, & 1 < \nu \leq 2, \end{cases},$$

$$\gamma_1(X) = 0 \text{ при } \nu > 3,$$

$$\gamma_2(X) = \begin{cases} \frac{6}{\nu-4}, & \nu > 4, \\ \infty, & 2 < \nu \leq 4. \end{cases}.$$

- Пусть  $Z \sim N(0, 1)$ ,  $V \sim \chi_{\nu}^2$ , тогда

$$T = \frac{Z}{\sqrt{V/\nu}} \sim St(\nu).$$

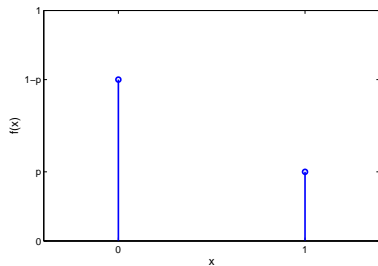
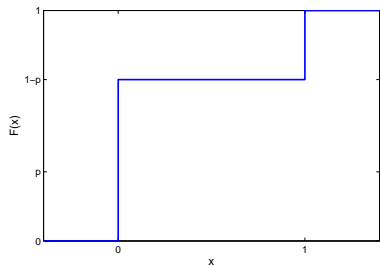
- Если  $X \sim St(\nu)$ , то

$$Y = \lim_{\nu \rightarrow \infty} X \sim N(0, 1).$$

# Распределение Бернулли

$X \in \{0, 1\} \sim \text{Ber}(p)$ ,  $p \in (0, 1)$  — распределение, моделирующее испытание Бернулли.

Пример: результат подбрасывания монеты.



$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - p, & 0 \leq x < 1, \\ 1, & x \geq 1, \end{cases}$$
$$f(x) = \begin{cases} 1 - p, & x = 0, \\ p, & x = 1. \end{cases}$$

## Распределение Бернулли

$$\mathbb{E}X = p,$$

$$\text{med } X = \begin{cases} 0, & 1 - p > p, \\ 0.5, & 1 - p = p, \\ 1, & 1 - p < p, \end{cases}$$

$$\text{mode } X = \begin{cases} 0, & 1 - p > p, \\ \{0, 1\}, & 1 - p = p, \\ 1, & 1 - p < p, \end{cases}$$

$$\mathbb{D}X = p(1 - p),$$

$$\gamma_1(X) = \frac{1 - 2p}{\sqrt{p(1 - p)}},$$

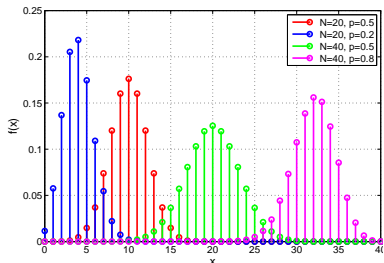
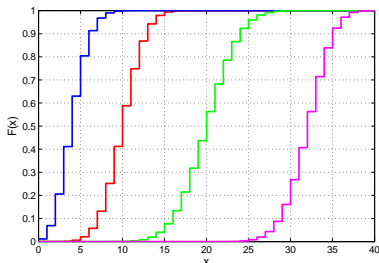
$$\gamma_2(X) = \frac{1 - 6p(1 - p)}{p(1 - p)}.$$



# Биномиальное распределение

$X \in \{0, \dots, N\} \sim \text{Bin}(N, p)$ ,  $N \in \mathbb{N}$ ,  $p \in [0, 1]$  — распределение числа успехов в  $N$  независимых испытаниях Бернулли.

Пример: число выпавших решек при независимом подбрасывании  $N$  монет.



$$F(x) = I_{1-p}(N-x, 1+x),$$

$$f(x) = C_N^x p^x (1-p)^{N-x}.$$

## Биномиальное распределение

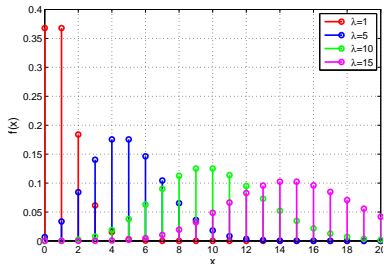
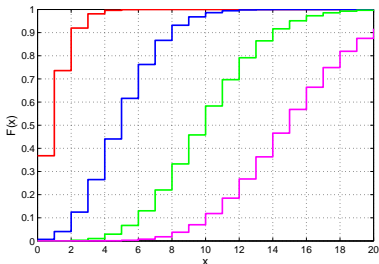
$$\begin{aligned}\mathbb{E}X &= Np, \\ \text{med } X &= \lfloor Np \rfloor \text{ или } \lceil Np \rceil, \\ \text{mode } X &= \lfloor (N+1)p \rfloor \text{ или } \lceil (N+1)p \rceil - 1, \\ \mathbb{D}X &= Np(1-p), \\ \gamma_1(X) &= \frac{1-2p}{\sqrt{Np(1-p)}}, \\ \gamma_2(X) &= \frac{1-Np(1-p)}{Np(1-p)}.\end{aligned}$$

- $X \sim \text{Bin}(1, p) \Leftrightarrow X \sim \text{Ber}(p)$ .
- Если  $N > 20$  и  $p$  не слишком близко к нулю или единице, то для  $X \sim \text{Bin}(N, p)$  справедливо приближение

$$F_X(x) \approx \Phi\left(\frac{x - Np}{\sqrt{Np(1-p)}}\right).$$

# Распределение Пуассона

$X \in \{0, 1, 2, \dots\} \sim Pois(\lambda)$ ,  $\lambda > 0$  — распределение числа независимых событий в фиксированном временном или пространственном интервале.



$$F(x) = e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!},$$
$$f(x) = e^{-\lambda} \frac{\lambda^x}{x!}.$$

## Распределение Пуассона

$$\begin{aligned}\mathbb{E}X &= \lambda, \\ \text{mode } X &= \lfloor \lambda \rfloor, \lceil \lambda \rceil - 1, \\ \mathbb{D}X &= \lambda, \\ \gamma_1(X) &= \lambda^{-1/2}, \\ \gamma_2(X) &= \lambda^{-1}.\end{aligned}$$

- Пусть  $X_1, \dots, X_n$  независимы,  $X_i \sim \text{Pois}(\lambda_i)$ , тогда

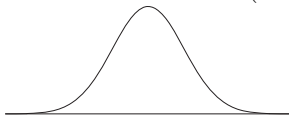
$$\sum_{i=1}^n X_i \sim \text{Pois} \left( \sum_{i=1}^n \lambda_i \right).$$

- Если  $X \sim \text{Pois}(\lambda)$ ,  $Y = \sqrt{X}$ , то

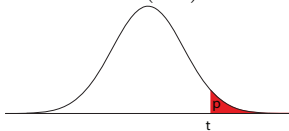
$$F_Y(x) \approx \Phi \left( \frac{x - \sqrt{\lambda}}{1/2} \right).$$

# Проверка гипотез

- выборка:  $X^n = (X_1, \dots, X_n), X \sim \mathbf{P} \in \Omega;$
- нулевая гипотеза:  $H_0: \mathbf{P} \in \omega, \omega \in \Omega;$
- альтернатива:  $H_1: \mathbf{P} \notin \omega;$
- статистика:  $T(X^n), T(X^n) \sim F(x)$  при  $\mathbf{P} \in \omega;$   
 $T(X^n) \not\sim F(x)$  при  $\mathbf{P} \notin \omega;$



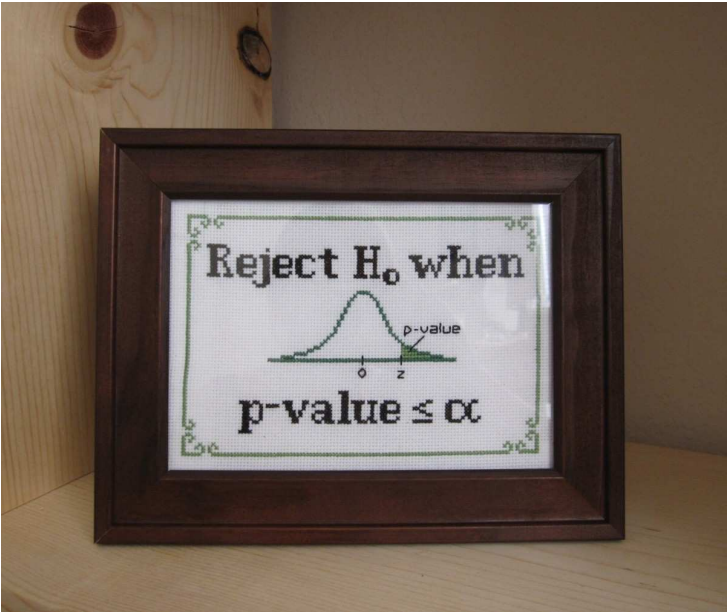
- реализация выборки:  $x^n = (x_1, \dots, x_n);$
- реализация статистики:  $t = T(x^n);$
- достижимый уровень значимости:  $p(x^n)$  — вероятность при  $H_0$  получить  $T(X^n) = t$  или ещё более экстремальное;



$$p(x^n) = \mathbf{P}(T \geq t | H_0)$$

Гипотеза отвергается при  $p(x^n) \leq \alpha, \alpha$  — уровень значимости.

# Проверка гипотез



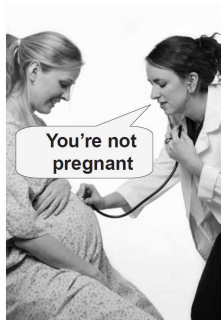
# Ошибки I и II рода

	$H_0$ верна	$H_0$ неверна
$H_0$ принимается	$H_0$ верно принята	Ошибка второго рода (False negative)
$H_0$ отвергается	Ошибка первого рода (False positive)	$H_0$ верно отвергнута

**Type I error**  
(false positive)



**Type II error**  
(false negative)



## Ошибки I и II рода

Задача проверки гипотез несимметрична относительно пары  $(H_0, H_1)$ : вероятность ошибки первого рода ограничивается малой величиной  $\alpha$ , второго рода — минимизируется путём выбора критерия.

**Мощность:**  $\text{pow} = \mathbf{P}(p(T) \leq \alpha | H_1)$ .

**Состоятельный критерий:**  $\text{pow} \rightarrow 1$  для всех альтернатив  $H_1$  при  $n \rightarrow \infty$ .

$T_1$  — равномерно наиболее мощный критерий, если  $\forall T_2$

$$\mathbf{P}(p(T_1) \leq \alpha | H_1) \geq \mathbf{P}(p(T_2) \leq \alpha | H_1) \quad \forall H_1 \neq H_0,$$

$$\mathbf{P}(p(T_1) \leq \alpha | H_0) = \mathbf{P}(p(T_2) \leq \alpha | H_0),$$

причём хотя бы для одной  $H_1$  неравенство строгое.



## Интерпретация результата

Если величина  $p$  достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Если величина  $p$  недостаточно мала, то данные не свидетельствуют против нулевой гипотезы в пользу альтернативы.

При помощи инструмента проверки гипотез нельзя доказать справедливость нулевой гипотезы!

Absence of evidence  $\nRightarrow$  evidence of absence.



## Статистическая и практическая значимость

Вероятность отвергнуть нулевую гипотезу зависят не только от того, насколько она отличается от истины, но и от размера выборки.

По мере увеличения  $n$  нулевая гипотеза может сначала приниматься, но потом выявятся более тонкие несоответствия выборки гипотезе  $H_0$ , и она будет отвергнута.

При любой проверке гипотез нужно считать **размер эффекта** — величину, оценивающую по данным, насколько нулевая гипотезу отличается от истины, и оценивать его практическую значимость.

## Статистическая и практическая значимость

- (Lee et al, 2010): за три года женщины, упражнявшиеся не меньше часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ( $p < 0.001$ ). Разница в набранном весе составила 150 г. Практическая значимость такого эффекта сомнительна. Подробности: <http://youtu.be/oqDZO-mfN4Q>.
- (Ellis, 2010, гл. 2): в 2002 году клинические испытания гормонального препарата Премарин, облегчающего симптомы менопаузы, были досрочно прерваны. Было обнаружено, что его приём ведёт к значимому увеличению риска развития рака груди на 0.08%, риска инсульта на 0.08% и инфаркта на 0.07%. Формально эффект крайне мал, но с учётом численности населения он превращается в тысячи дополнительных смертей.
- (Kirk, 1996): если при испытании гипотетического лекарства, позволяющего замедлить прогресс ослабления интеллекта больных Альцгеймером оказывается, что разница в IQ контрольной и тестовой групп составляет 13 пунктов, возможно, изучение лекарства стоит продолжить, даже если эта разница статистически незначима.

## Другие особенности

- Выбранная статистика может отражать не всю информацию, содержащуюся в выборке. Пример:

$$H_0: X \sim N(\mu, \sigma^2), \quad H_1: H_0 \text{ неверна};$$

$$T(X^n) = g_1.$$

Все симметричные распределения будут признаны нормальными!

- Гипотезы вида  $H_0: \theta = \theta_0$  можно проверять при помощи доверительных интервалов для  $\theta$ : если  $\theta_0$  не попадает в  $100(1 - \alpha)\%$  доверительный интервал для  $\theta$ , то  $H_0$  отвергается на уровне значимости  $\alpha$ .

# Shaken, not stirred

Джеймс Бонд говорит, что предпочитает мартини смешанным, но не взболтанным. Проведём слепой тест:  $n$  раз предложим ему пару напитков и выясним, какой из двух он предпочитает.

Выборка: бинарный вектор длины  $n$ , 1 — Джеймс Бонд предпочёт смешанный, 0 — взболтанный.

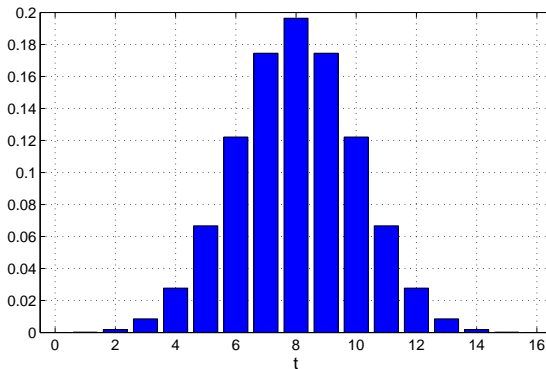
Нулевая гипотеза: Джеймс Бонд не различает два вида мартини, т. е., выбирает наугад.

Статистика  $t$  — число единиц в выборке.

# Нулевое распределение

Если нулевая гипотеза справедлива и Джеймс Бонд не различает два вида мартины, то равновероятны все выборки длины  $n$  из нулей и единиц.

Пусть  $n = 16$ , тогда существует  $2^{16} = 65536$  равновероятных варианта. Статистика  $t$  принимает значения от 0 до 16:

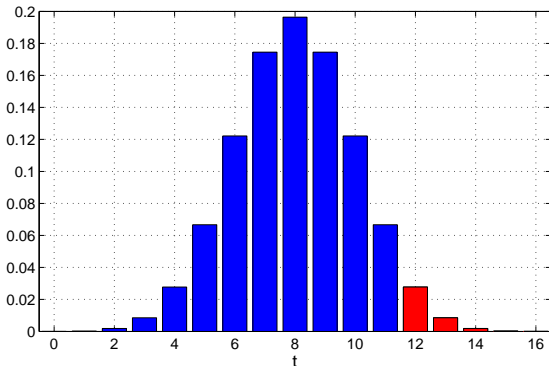


# Односторонняя альтернатива

$H_1$ : Джеймс Бонд предпочитает смешанный мартини.

При справедливости такой альтернативы более вероятны большие значения  $t$  (т.е., большие  $t$  свидетельствуют против  $H_0$  в пользу  $H_1$ ).

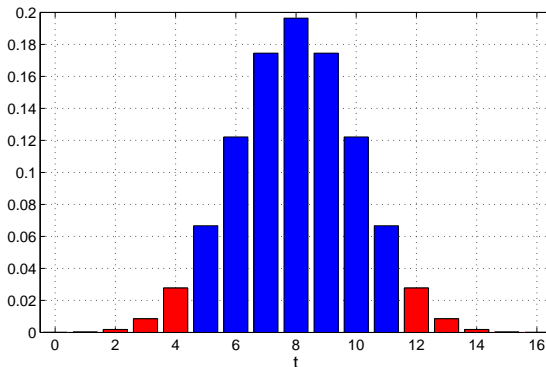
Вероятность того, что Джеймс Бонд предпочтёт смешанный мартини в 12 или более случаях из 16 при справедливости  $H_0$ , равна  $\frac{2517}{65536} \approx 0.0384$ .



0.0384 — достигаемый уровень значимости при реализации  $t = 12$ .

## Двусторонняя альтернатива

$H_1$ : Джеймс Бонд предпочитает какой-то определённый вид мартини. При справедливости такой альтернативы и очень большие, и очень маленькие значения  $t$  свидетельствуют против  $H_0$  в пользу  $H_1$ ). Вероятность того, что Джеймс Бонд предпочтёт смешанный мартини в 12 или более случаях из 16 при справедливости  $H_0$ , равна  $\frac{5034}{65536} \approx 0.0768$ .



0.0768 — достигаемый уровень значимости при реализации  $t = 12$ .



## Достижимый уровень значимости

Чем ниже достижимый уровень значимости, тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

0.0384 — вероятность реализации  $t \geq 12$  при условии, что нулевая гипотеза справедлива, т. е. Джеймс Бонд выбирает мартини наугад.

Достижимый уровень значимости нельзя интерпретировать как вероятность справедливости нулевой гипотезы!

## Достижимый уровень значимости

**Пример:** утверждается, что осьминог предсказывает результаты матчей с участием сборной Германии на чемпионате мира по футболу 2010 года, выбирая кормушку с флагом страны-победителя. По результатам 13 испытаний ему удаётся верно угадать результаты 11 матчей. Аналогичный предыдущему критерий даёт достижимый уровень значимости  $p \approx 0.0112$ .



0.0112 — не вероятность того, что осьминог выбирает кормушку наугад!  
Эта вероятность равна единице.

$$p = \mathbf{P}(T \geq t | H_0) \neq \mathbf{P}(H_0 | T \geq t).$$

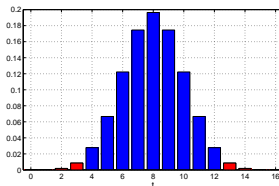
## Достижимый уровень значимости

**Пример:** пусть Джеймс Бонд выбирает смешанный martini в 51% случаев (ненаблюдаемая вероятность).

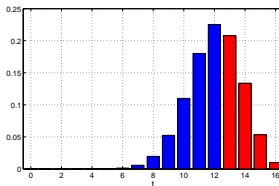
Пусть по итогам 100 испытаний смешанный martini был выбран 49 раз. Достижимый уровень значимости против односторонней альтернативы —  $p \approx 0.6178$ . Нулевая гипотеза не отвергается, при этом сказать, что она верна, было бы ошибкой — Джеймс Бонд выбирает смешанный и взболтанный martini не с одинаковыми вероятностями!

# Мощность

Проверяя нулевую гипотезу против двусторонней альтернативы, мы отвергаем  $H_0$  при  $t \geq 13$  или  $t \leq 3$ , что обеспечивает достигаемый уровень значимости  $p = 0.0213 \leq \alpha = 0.05$ .

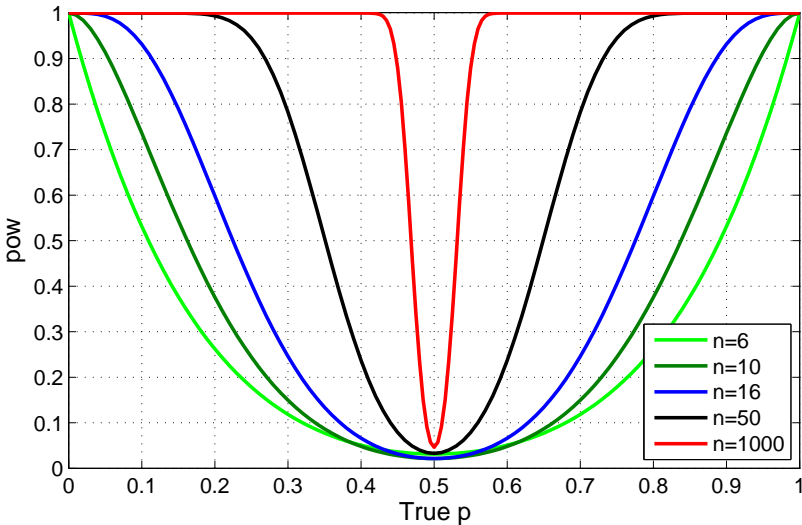


Пусть Джеймс Бонд выбирает смешанный мартини в 75% случаев.



$\text{pow} \approx 0.6202$ , т. е., при многократном повторении эксперимента гипотеза будет отклонена только в 62% случаев.

# Мощность



## Мощность

Мощность критерия зависит от следующих факторов:

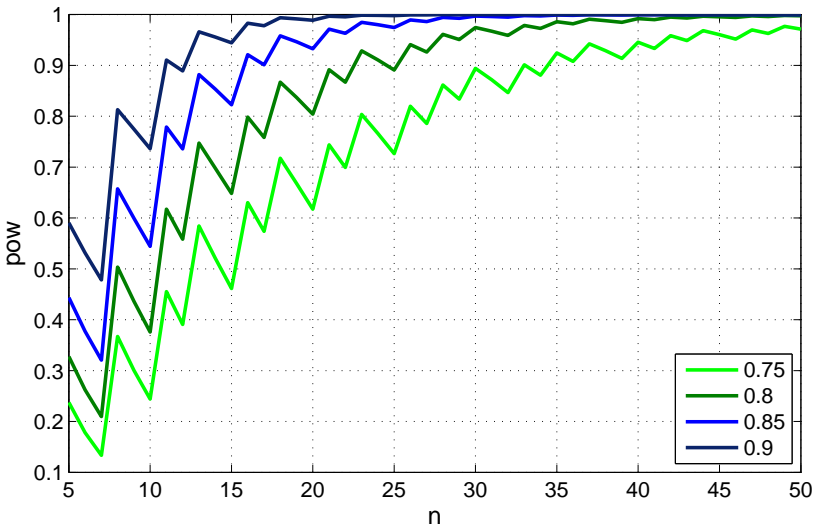
- размер выборки;
- размер отклонения от нулевой гипотезы;
- чувствительность статистики критерия;
- тип альтернативы.

## Размер выборки

Особенности прикладной задачи: 1 порция мартини содержит 55 мл джина и 15 мл вермута — суммарно около 25 мл спирта. Смертельная доза алкоголя при массе тела 80 кг составляет от 320 до 960 мл спирта в зависимости от толерантности (от 13 до 38 мартини).

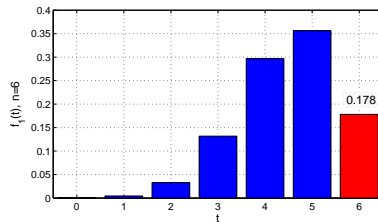
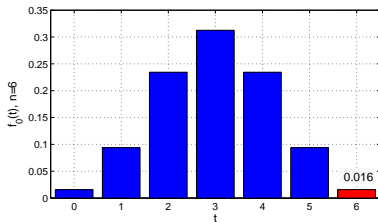
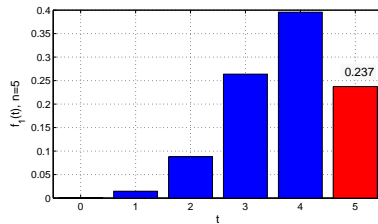
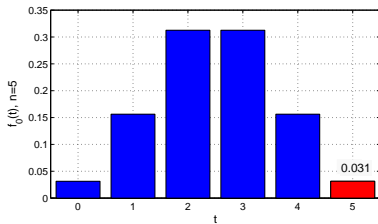
Обеспечение требуемой мощности: размеры выборки подбирается так, чтобы при размере отклонения от нулевой гипотезы не меньше заданного (например, вероятность выбора смешанного мартини не меньше 0.75) мощность была не меньше заданной.

# Шокирующий график





# Падение мощности: объяснение



## Литература

### Справочники по статистике:

- Кобзарь А.И. *Прикладная математическая статистика*. — М.: Физматлит, 2006.
- Kanji G.K. *100 statistical tests*. — London: SAGE Publications, 2006.

### Вводные учебники по статистике:

- Good P.I., Hardin J.W. *Common Errors in Statistics (and How to Avoid Them)*. — Hoboken: John Wiley & Sons, 2003.
- Reinhart A. *Statistics Done Wrong. The woefully complete guide*. — <http://www.statisticsonewrong.com/>

### R:

- [http://youtu.be/jwBgGS\\_4RQA](http://youtu.be/jwBgGS_4RQA)
- <http://cran.r-project.org/doc/contrib/Short-refcard.pdf>
- <http://adv-r.had.co.nz/>
- <https://www.coursera.org/course/rprog>

Ellis P.D. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. — Cambridge: Cambridge University Press, 2010.

Kirk R.E. (1996). *Practical Significance: A Concept Whose Time Has Come*. *Educational and Psychological Measurement*, 56(5), 746–759.

Lee I.-M., Djoussè L., Sesso H.D., Wang L., Buring J.E. (2010). *Physical Activity and Weight Gain Prevention*. *JAMA: the journal of the American Medical Association*, 303(12), 1173–1179.