Прикладная статистика 1. Введение.

12 февраля 2013 г.

Выборка

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

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

$$X^n = (X_1, X_2, \ldots, X_n)\,.$$

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

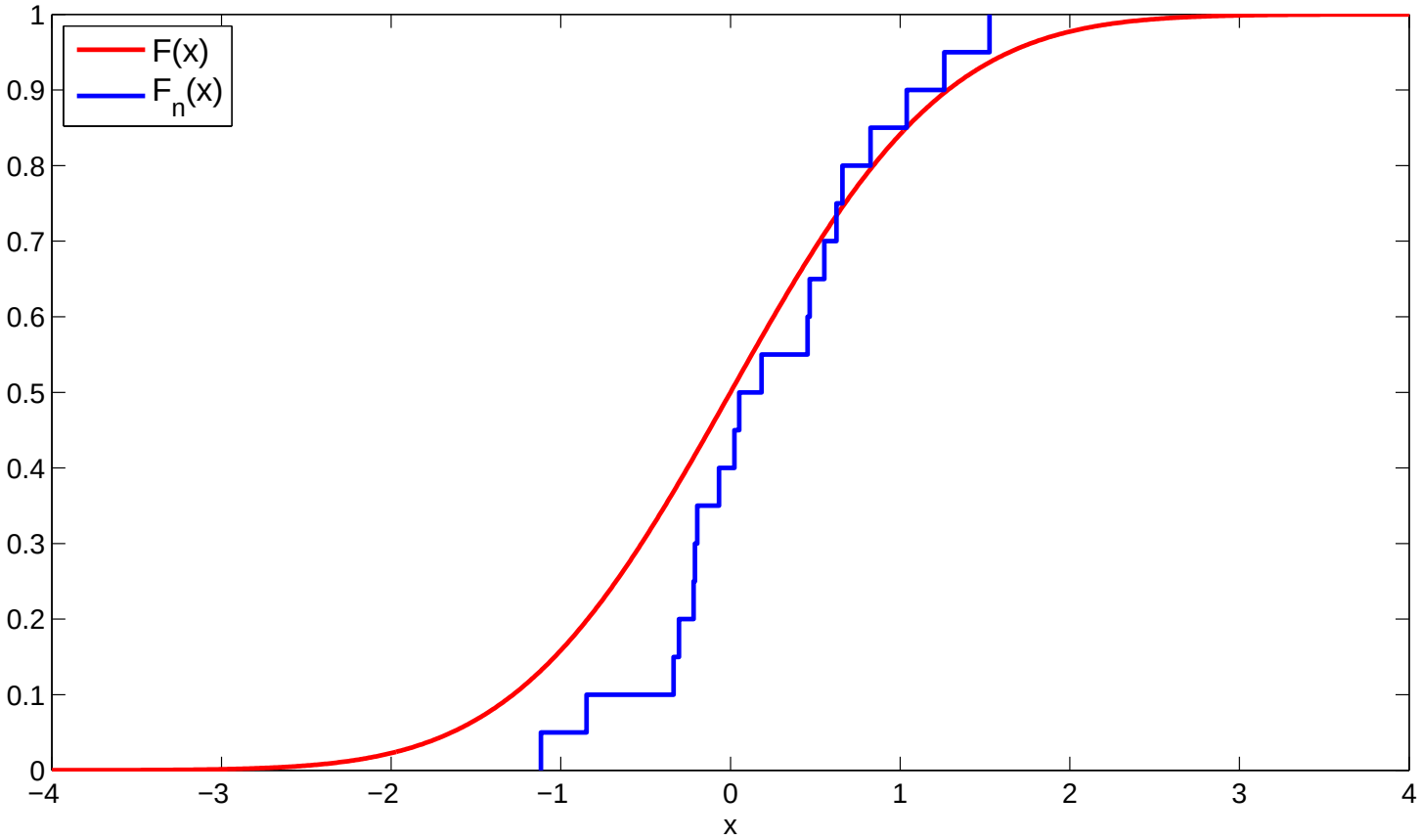
Пусть $F(x)$ — функция распределения элемента простой выборки:

$$F(x) = P\,(X_1 < x)\,.$$

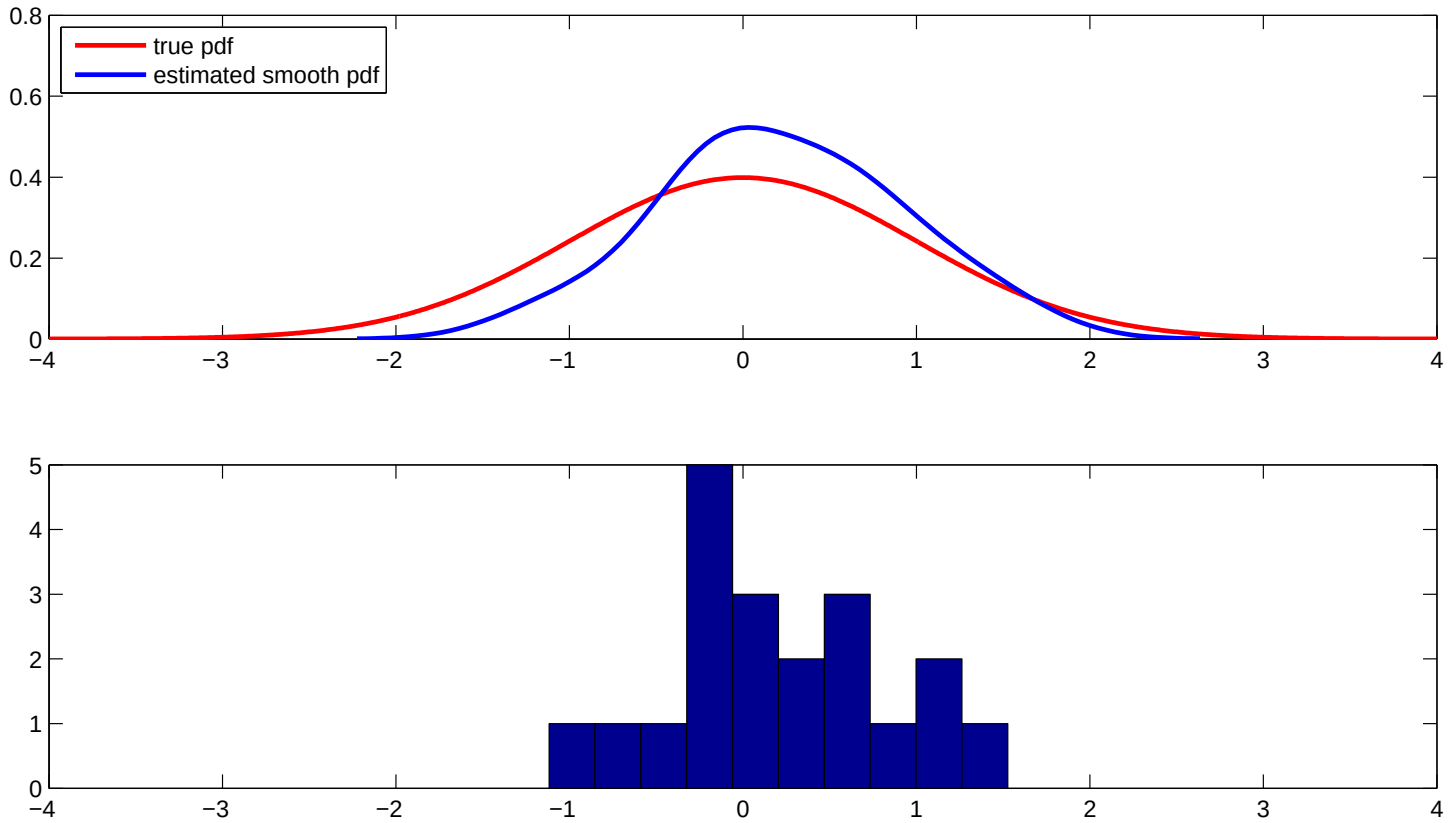
Основная задача статистики — описание $F(x)$ по реализации выборки.

Функция распределения

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



Плотность распределения



Статистика

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

Примеры:

- **выборочное среднее:**

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

- **выборочная дисперсия:**

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

- **несмещённая выборочная дисперсия:**

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

Статистика

Вариационный ряд:

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

- $X_{(k)}$ — k-я **порядковая статистика;**

Квантиль порядка α случайной величины X:

$$X_\alpha : P\left(X \leq X_\alpha\right) \geq \alpha, P\left(X \geq X_\alpha\right) \geq 1 - \alpha.$$

- **выборочный α-квантиль:** $X_{([n\alpha])}$;
- **выборочная медиана:**

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

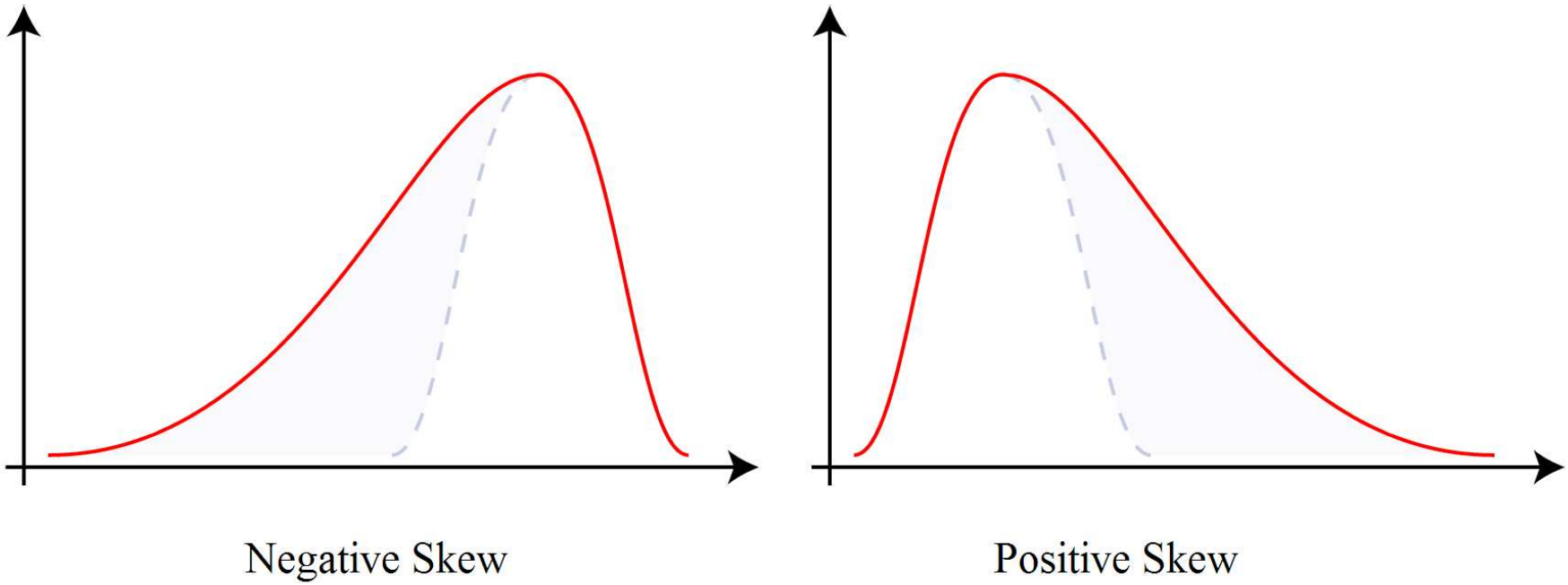
- **выборочный интерквартильный размах:**

$$IQR = X_{([3n/4])} - X_{([n/4])}.$$

Статистика

Коэффициент асимметрии (skewness):

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



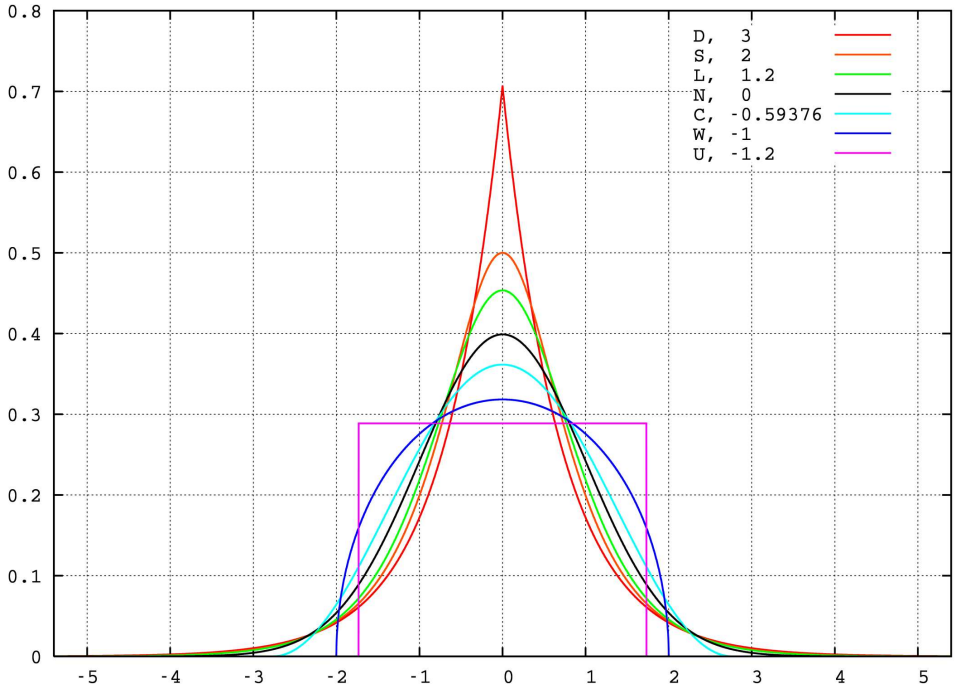
- **выборочный коэффициент асимметрии:**

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

Статистика

Коэффициент эксцесса (kurtosis):

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

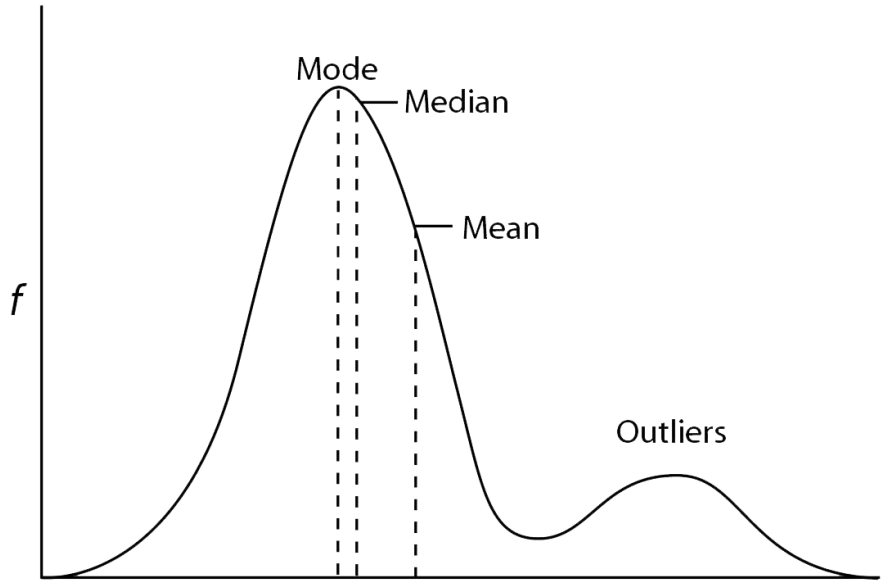


- **выборочный коэффициент эксцесса:**

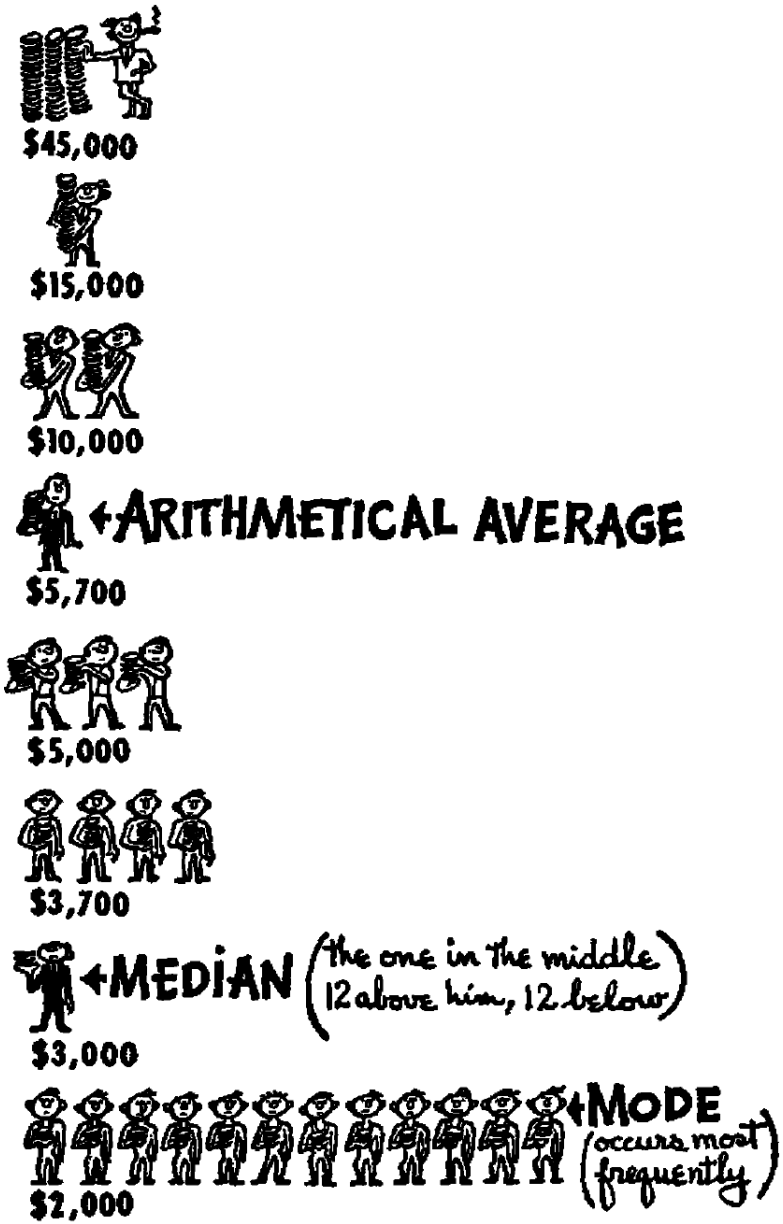
$$g_2 = \frac{n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^2} - 3.$$

Оценки центральной тенденции

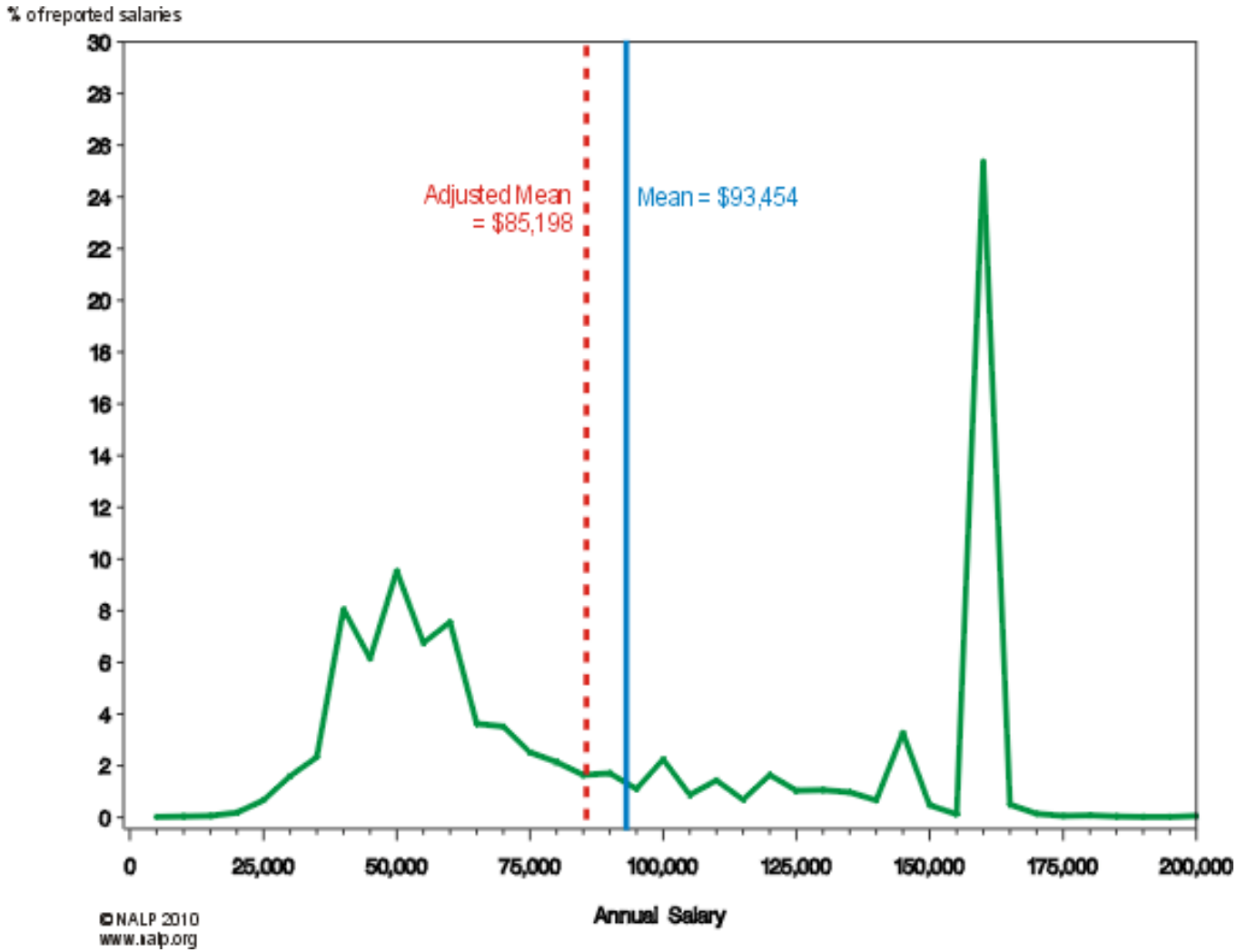
Выборочное среднее — среднее значение в выборке.
Медиана — средний элемент вариационного ряда выборки.
Мода — самое распространённое значение в выборке.



How to lie with statistics (Huff, 1954)



Данные NALP (The Association for Legal Career Professionals)



Salary Distribution Curve for the Class of 2009 Shows Relatively Few Salaries Were Close to the Mean

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: $\hat{\theta}_n \xrightarrow{P} \theta$ при $n \to \infty$.
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: $\hat{\theta}_n$ устойчива относительно малых отклонений реального распределения X от $F(x, \theta)$ (в частности, относительно выбросов).

Интервальные оценки

Оценим параметр θ двумя статистиками:

$$P\left(\theta \notin \left[\hat{\theta}_{1n}, \hat{\theta}_{2n}\right]\right) \leq \alpha.$$

α — уровень доверия;
$\hat{\theta}_{1n}, \hat{\theta}_{2n}$ — 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 параметра.

Интервальные оценки

Пример: доверительный интервал для среднего $X \sim N(\theta, \sigma^2)$ при известном σ.

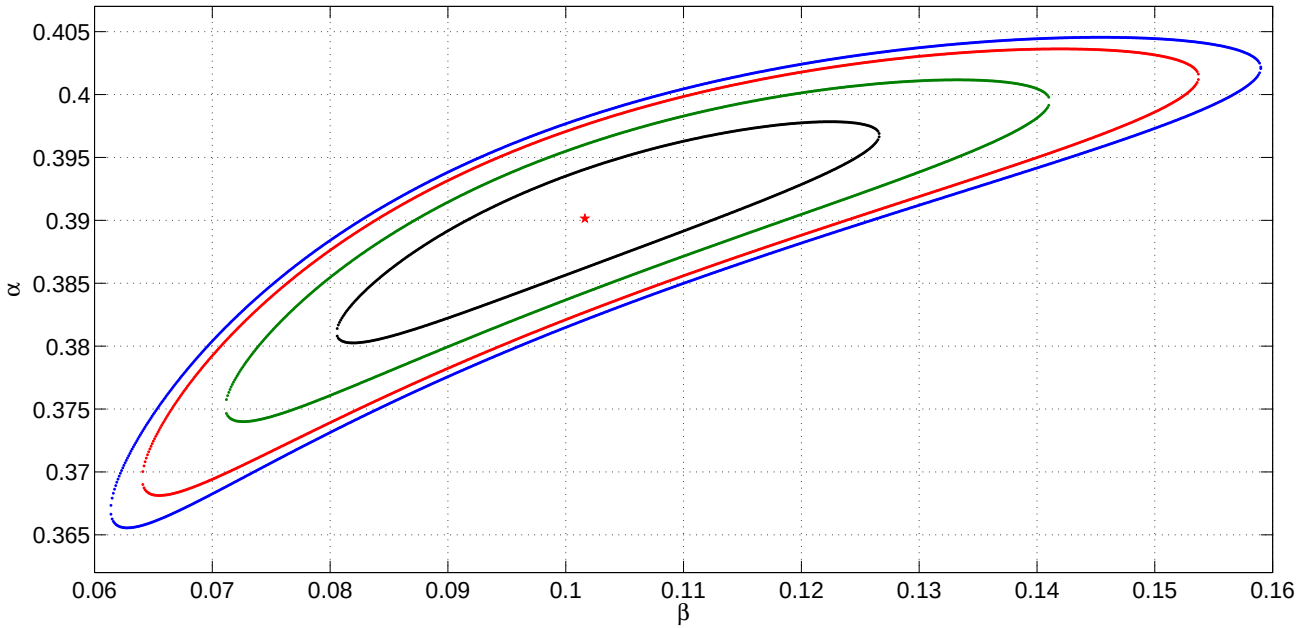
$$\bar{X} \sim N\left(\theta, \frac{\sigma^2}{n}\right) \Rightarrow \sqrt{n}\frac{\bar{X} - \theta}{\sigma} \sim N(0,1) \Rightarrow$$

$$\hat{\theta}_{1n} = \bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\theta}_{2n} = \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Правило двух сигм: если $X \sim N\left(\mu, \sigma^2\right)$, то $P\left(|X - \mu| \leq 2\sigma\right) \approx 0.954$. Если X распределена не нормально, то можно утверждать только $P\left(|X - \mathbb{E}X| \leq 2\sqrt{\mathbb{D}X}\right) \geq \frac{3}{4}$ (из неравенства Чебышева).

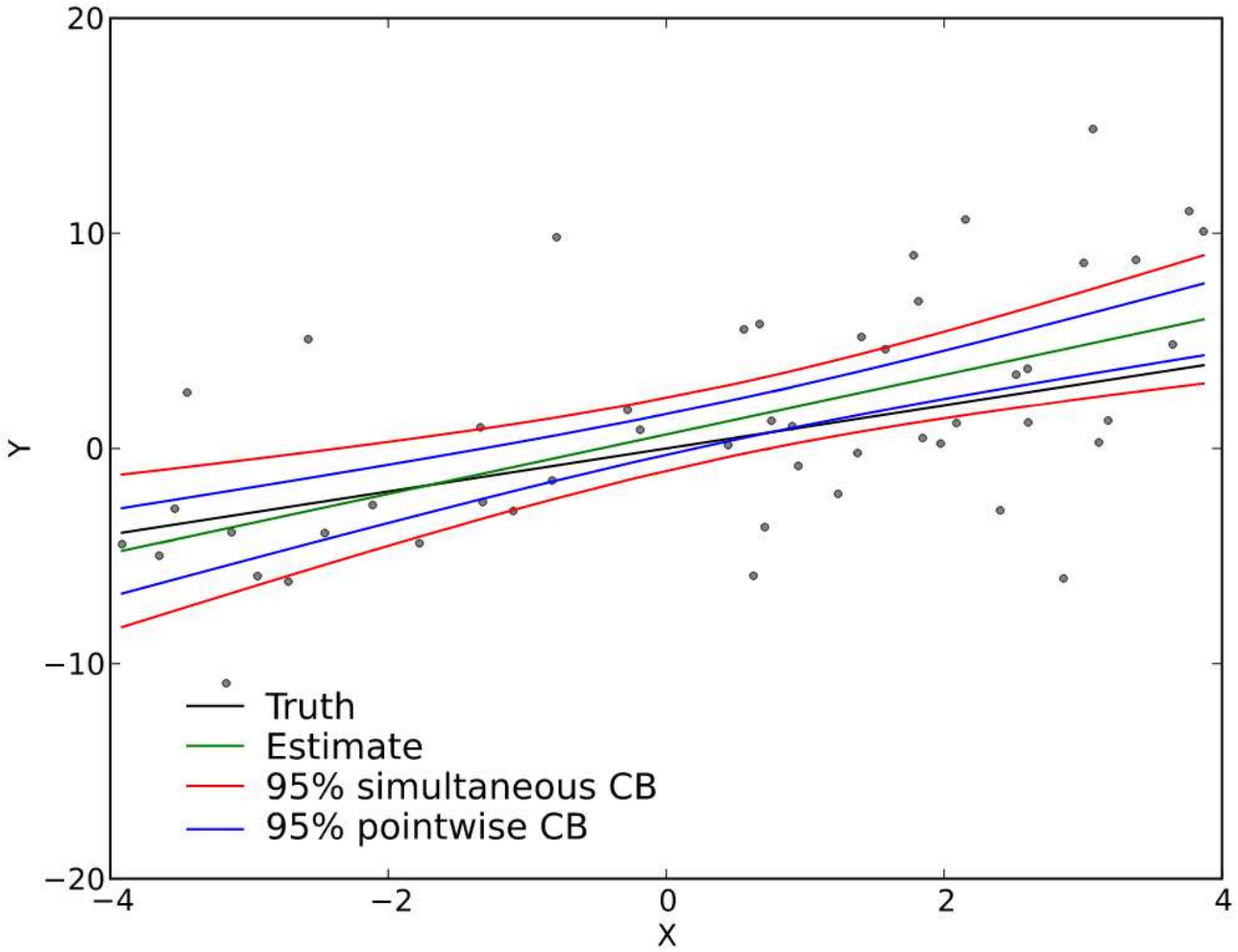
Интервальные оценки

Доверительная область для пары неизвестных параметров (α, β):



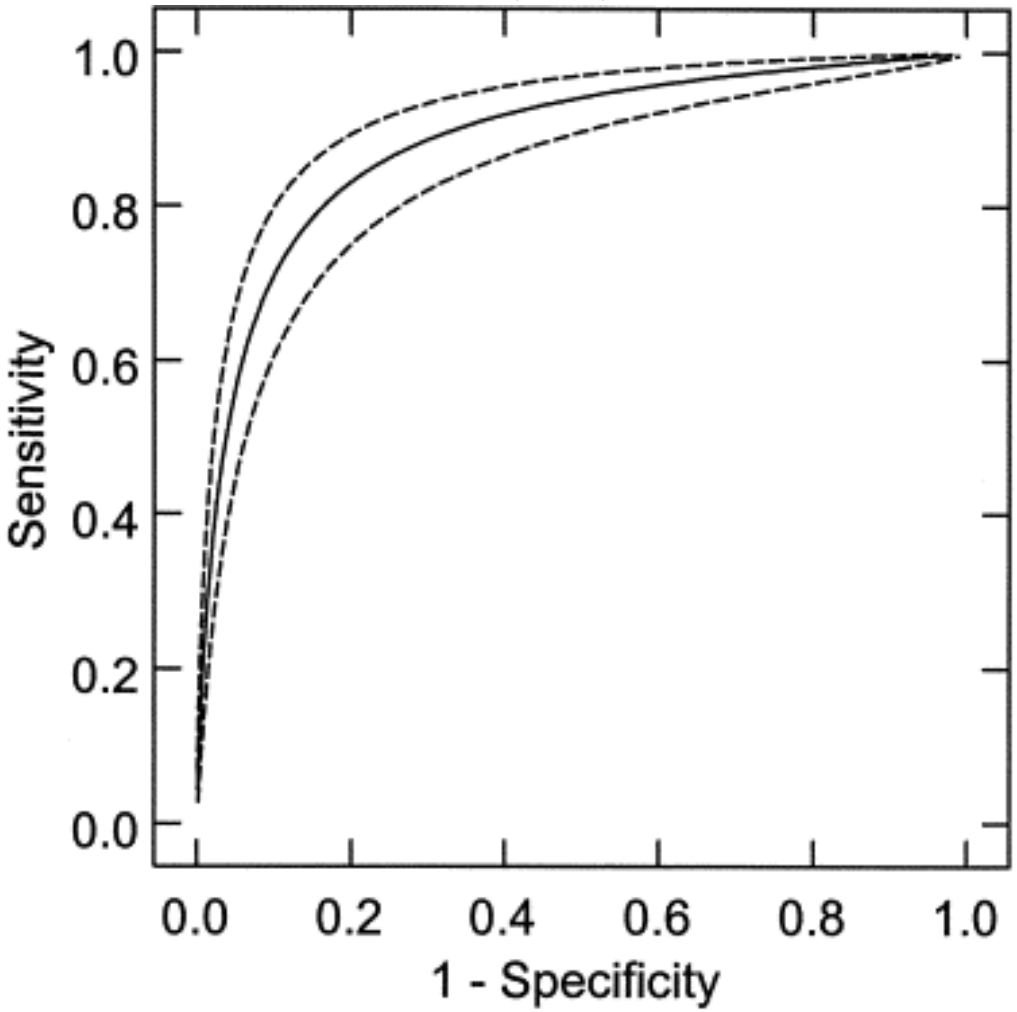
Интервальные оценки

Доверительная лента для функции $Y = f(x)$:

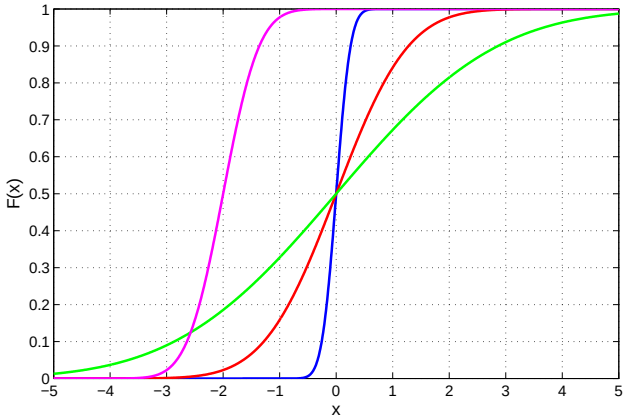


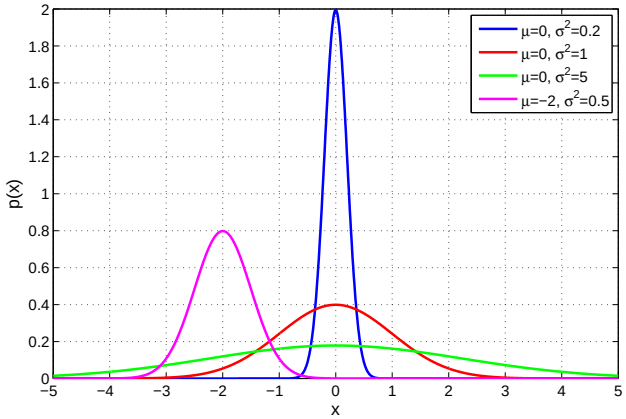
Интервальные оценки

Доверительная лента для функции $f(x, y)$:



Нормальное распределение





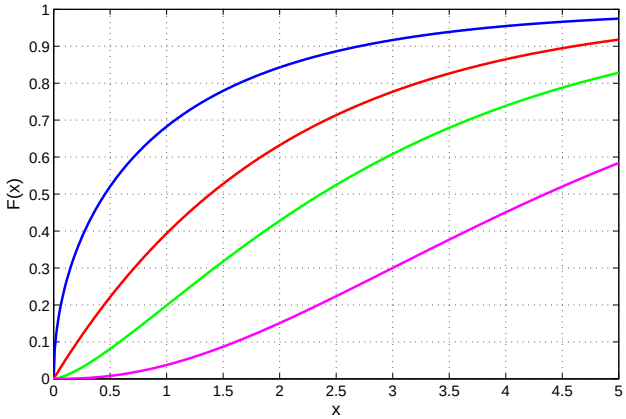
$$X \sim N\left(\mu, \sigma^2\right), \sigma^2 > 0$$

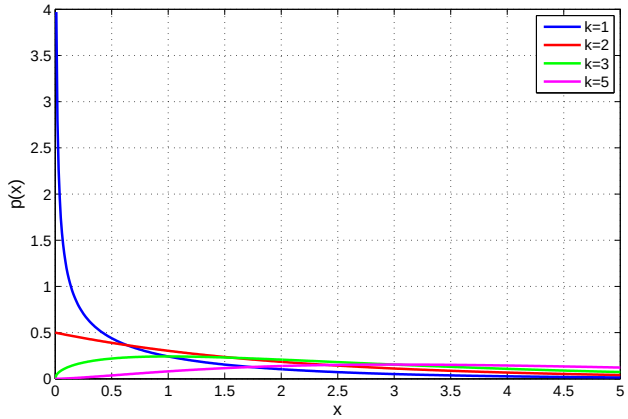
$$F(x) = \frac{1}{2}\left(1 + \Phi\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right), \quad p\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}X = medX = modeX = \mu, \quad \mathbb{D}X = \sigma^2$$

$$\gamma_1(X) = 0, \quad \gamma_2(X) = 0$$

Распределение хи-квадрат





$$X \sim \chi^2_k, k \in \mathbb{N}$$

$$F(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)}\gamma\left(\frac{k}{2}, \frac{x}{2}\right), \quad p(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$$

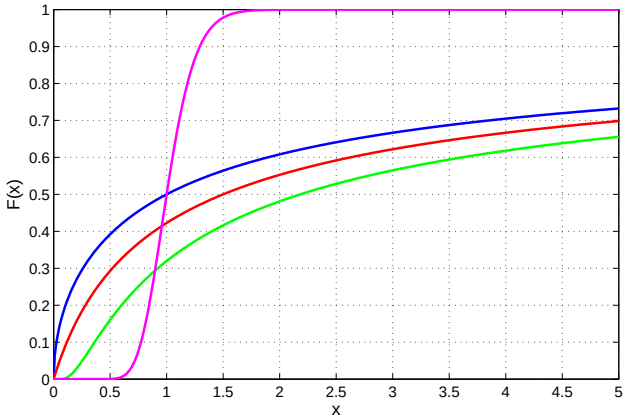
$$\mathbb{E}X = k, medX \approx k\left(1 - \frac{2}{9k}\right)^3, \quad modeX = \max(k-2, 0), \quad \mathbb{D}X = 2k$$

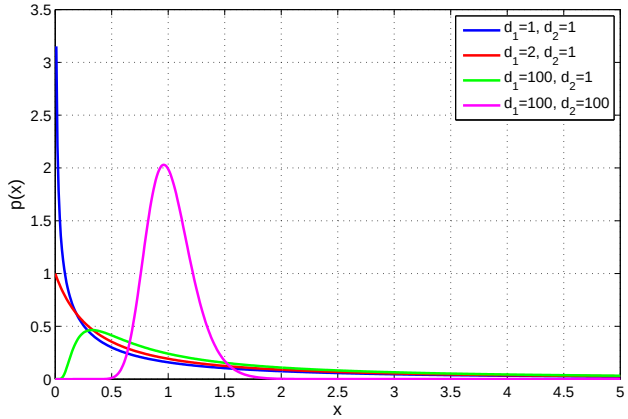
$$\gamma_1(X) = \sqrt{8/k}, \quad \gamma_2(X) = 12/k$$

Пусть $X_1, \ldots, X_k$ i.i.d., $X \sim N(0,1)$, тогда

$$\sum_{i=1}^{k} X_i^2 \sim \chi^2_k$$

Распределение Фишера





$$X \sim F\left(d_1, d_2\right), d_1, d_2 > 0$$

$$F(x) = I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right), \quad p\left(x\right) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

$$\mathbb{E}X = \frac{d_2}{d_2 - 2} \text{ при } d_2 > 2, \quad modeX = \frac{d_1 - 2}{d_1}\frac{d_2}{d_2 + 2} \text{ при } d_1 > 2$$

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

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

Распределение Фишера

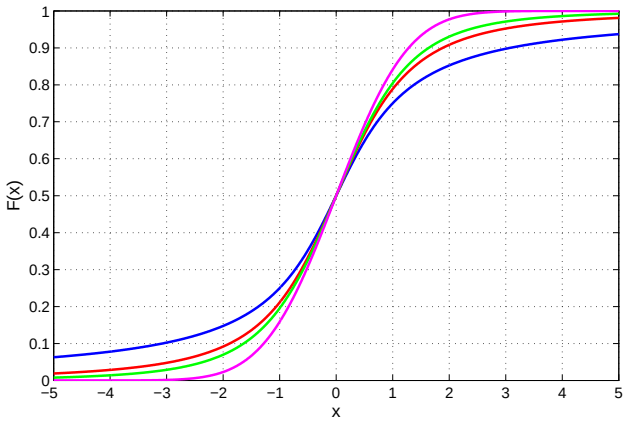
Пусть $X_1 \sim \chi^2_{d_1}, X_2 \sim \chi^2_{d_2}$, X_1 и X_2 независимы, тогда

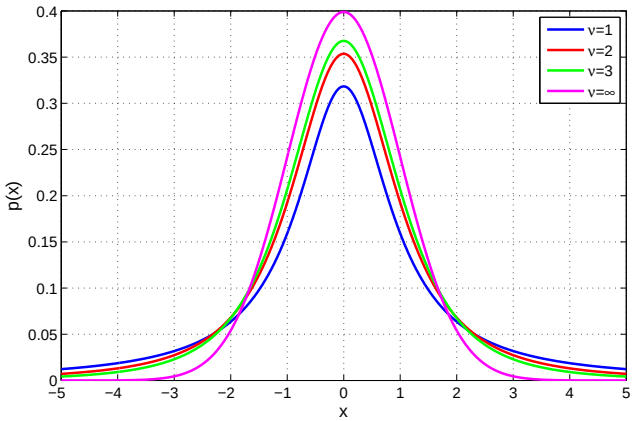
$$\frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2)$$

Если $X \sim F(d_1, d_2)$, то

$$Y = \lim_{d_2 \to \infty} d_1 X \sim \chi^2_{d_1}$$

Распределение Стьюдента





$$X \sim St(\nu), \nu > 0$$

$$F(x) = \frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right), \quad p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\mathbb{E}X = medX = modeX = 0 \text{ при } \nu > 1, \quad \mathbb{D}X = \frac{\nu}{\nu-2} \text{ при } \nu > 2, \infty \text{ при } 1 < \nu \leq 2$$

$$\gamma_1(X) = 0 \text{ при } \nu > 3, \quad \gamma_2(X) = \frac{6}{\nu-4} \text{ при } \nu > 4, \infty \text{ при } 2 < \nu \leq 4$$

Пусть $T \sim St(\nu), Z \sim N(0,1), V \sim \chi^2_\nu$, тогда

$$T = \frac{Z}{\sqrt{V/\nu}}$$

Распределение Бернулли

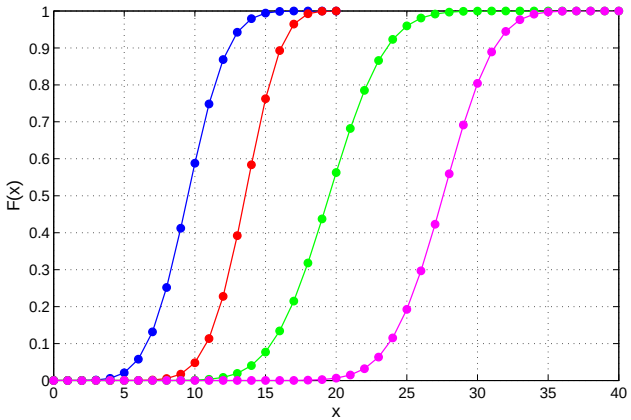
$$X \sim Bern\left(p\right), p \in (0,1)$$

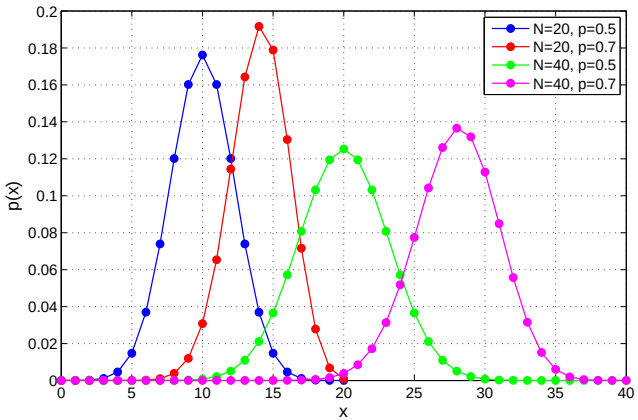
$$F(x) = \begin{cases} 0, & x < 0, \\ 1-p, & 0 \leq x < 1, , \\ 1, & x \geq 1, \end{cases} \quad p\left(x\right) = \begin{cases} 1-p, & x = 0, \\ p, & x = 1 \end{cases}$$

$$\mathbb{E}X = p, \quad medX = \begin{cases} 0, & 1-p > p, \\ 0.5, & 1-p = p, , \\ 1, & 1-p < p, \end{cases} \quad modeX = \begin{cases} 0, & 1-p > p, \\ 0,1, & 1-p = p, \\ 1, & 1-p < p, \end{cases}$$

$$\mathbb{D}X = p(1-p), \quad \gamma_1(X) = \frac{1-2p}{\sqrt{p(1-p)}}, \quad \gamma_2(X) = \frac{1-6p(1-p)}{p(1-p)}$$

Биномиальное распределение





$$X \sim Bin\left(N,p\right), N \in \mathbb{N}_0, p \in [0,1]$$

$$F(x) = I_{1-p}(N-x, 1+x), \quad p\left(x\right) = C_N^x p^x (1-p)^{N-x}$$

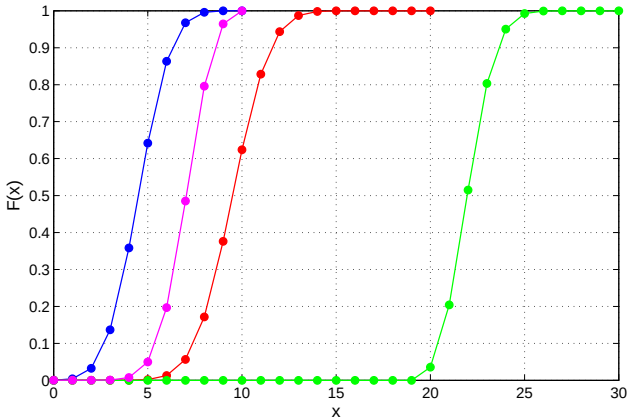
$$\mathbb{E}X = Np, \quad medX = \lfloor Np \rfloor \text{ или } \lceil Np \rceil, \quad modeX = \lfloor (N+1)p \rfloor \text{ или } \lceil (N+1)p \rceil - 1$$

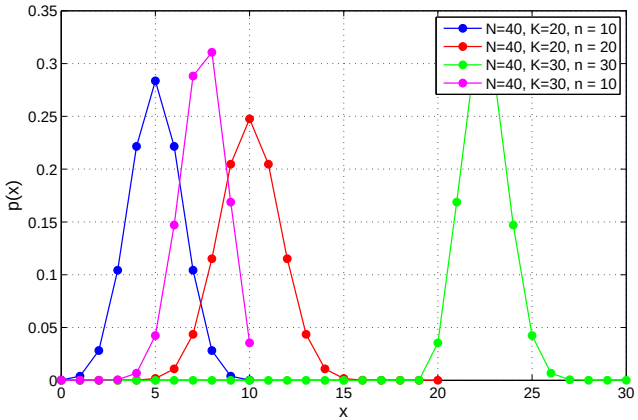
$$\mathbb{D}X = Np(1-p), \quad \gamma_1(X) = \frac{1-2p}{\sqrt{Np(1-p)}}, \quad \gamma_2(X) = \frac{1-Np(1-p)}{Np(1-p)}$$

$X \sim Bin(1,p)$ равносильно $X \sim Bern(p)$

Если $N > 20$ и p не слишком близко к нулю или единице, то для $X \sim Bin\left(N,p\right)$ справедливо $X \sim\approx N\left(Np, Np(1-p)\right)$.

Гипергеометрическое распределение





$$X \sim Hyp(K, N, n), N \in \mathbb{N}_0, K \in \{0, 1, \ldots, N\}, n \in \{0, 1, \ldots, N\}$$

$$X \in \{\max(0, n + K - N), \ldots, \min(K, n)\}$$

$$p(x) = \frac{C_K^x C_{N-K}^{n-x}}{C_N^n}$$

$$\mathbb{E}X = n\frac{K}{N}, \quad modeX = \left\lfloor \frac{(n+1)(K+1)}{N+2} \right\rfloor, \quad \mathbb{D}X = n\frac{K}{N}\frac{(N-K)}{N}\frac{N-n}{N-1}$$

$$\gamma_1(X) = \frac{(N-2K)(N-1)^{\frac{1}{2}}(N-2n)}{[nK(N-K)(N-n)]^{\frac{1}{2}}(N-2)}$$

Гипергеометрическое распределение

$X \sim Hyp(K, N, 1)$ равносильно $X \sim Bern(\frac{K}{N})$
Пусть $X \sim Hyp(K, N, n), Y \sim Bin(n, \frac{K}{N})$; если $\frac{K}{N}$ не близко к нулю или единице, а N и K велики по сравнению с n и $\frac{K}{N}$, то

$$P(X \leq k) \approx P(Y \leq k)$$

Пусть $X \sim Hyp(K, N, n)$; если n велико, $\frac{K}{N}$ не близко к нулю или единице, а N и K велики по сравнению с n и $\frac{K}{N}$, то

$$P(X \leq k) \approx \Phi\left(\frac{k - np}{\sqrt{np(1 - np)}}\right)$$

Проверка гипотез

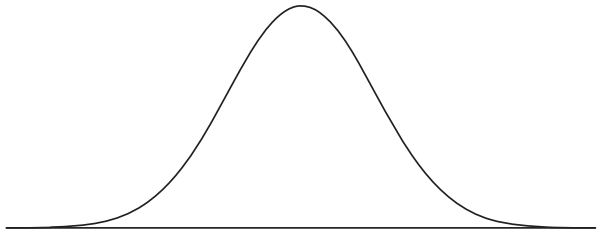
выборка: $X^n = \{X_1, \ldots, X_n\} \sim P \in \Omega;$

нулевая гипотеза: $H_0 \colon P \in \omega, \ \ \omega \in \Omega;$

альтернатива: $H_1 \colon P \notin \omega;$

статистика: $T(X^n), \ \ T(X^n) \sim F(x)$ при $P \in \omega;$

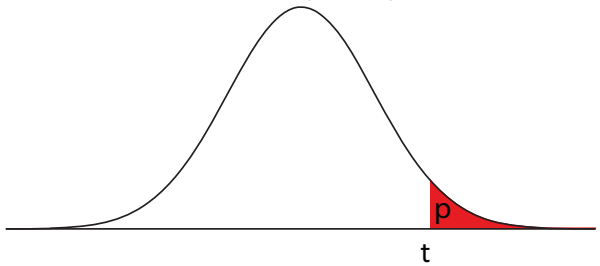
$T(X^n) \not\sim F(x)$ при $P \notin \omega;$



реализация выборки: $x^n = \{x_1, \ldots, x_n\};$

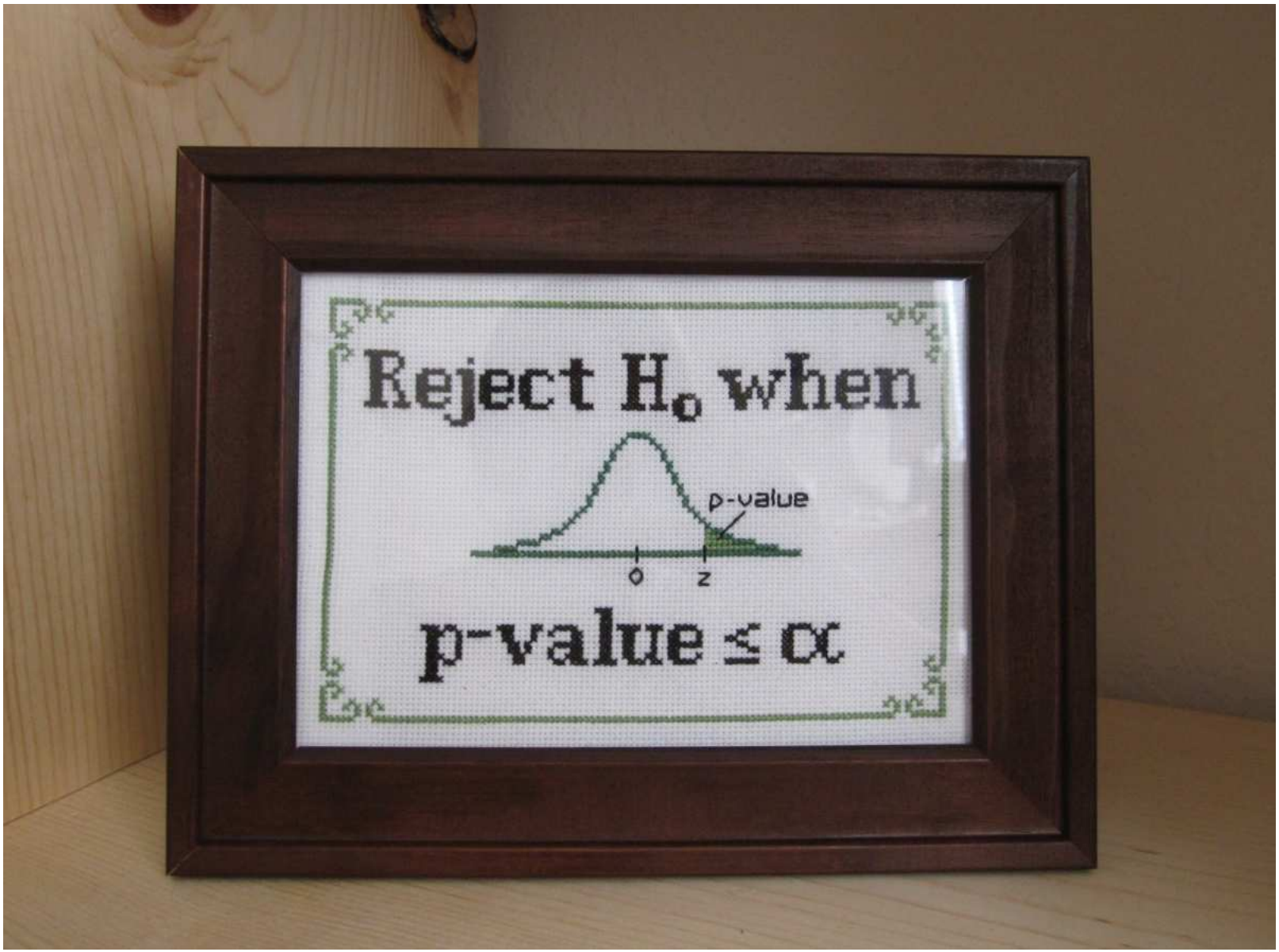
реализация статистики: $t = T(x^n);$

достигаемый уровень значимости: $p(T)$— вероятность при H_0 получить $T(X^n) = t$ или ещё более экстремальное;



Гипотеза отвергается при $p(t) \leq \alpha, \ \ \alpha$— уровень значимости.

Проверка гипотез



Ошибки I и II рода

Задача проверки гипотез несимметрична относительно пары (H_0, H_1).

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода
H_0 отвергается	Ошибка первого рода	H_0 верно отвергнута

Вероятность ошибки первого рода жёстко ограничивается малой величиной α — H_0 отвергается при $p \leq \alpha$.
Вероятность ошибки второго рода минимизируется путём выбора достаточно мощного критерия.
Мощность: $pow = P\left(p\left(T\right) \leq \alpha \,|H_1\right).$
Состоятельный критерий: $pow \to 1$ для всех альтернатив H_1 при $n \to \infty$.
Равномерно наиболее мощный критерий:

$$P\left(p\left(T_1\right) \leq \alpha \,|H_1\right) \leq P\left(p\left(T_2\right) \leq \alpha \,|H_1\right) \forall H_1 \neq H_0,$$

$$P\left(p\left(T_1\right) \leq \alpha \,|H_0\right) = P\left(p\left(T_2\right) \leq \alpha \,|H_0\right),$$

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 $\not\Rightarrow$ evidence of absence.

Другие особенности

- По мере увеличения n нулевая гипотеза может сначала приниматься, но потом выявятся более тонкие несоответствия выборки H_0, и она будет отвергнута.
- Выбранная статистика может отражать не всю априорную информацию, содержащуюся в H_0.

Пример:

$$H_0 : X \sim N\left(\mu, \sigma^2\right),$$

$$T\left(X^n\right) = g_1.$$

Все симметричные распределения будут признаны нормальными!

- Гипотезы вида $H_0 : \theta = \theta_0$ можно проверять при помощи доверительных интервалов для θ: если $\theta_0 \in \left[\hat{\theta}_{1n}, \hat{\theta}_{2n}\right]$, нулевая гипотеза не отвергается.

Shaken, not stirred

Джеймс Бонд говорил, что предпочитает мартини смешанным, но не взболтанным.

Слепой тест: n раз предложим Джеймсу Бонду пару напитков и выясним, какой из двух он предпочитает.

Выборка: бинарный вектор длины n, “1” — Джеймс Бонд предпочёл смешанный, “0” — взболтанный.

Нулевая гипотеза H_0: Джеймс Бонд не различает два вида мартини, т. е., выбирает наугад.

Статистика t — число единиц в выборке, т. е., число испытаний, в которых Джеймс Бонд предпочёл смешанный мартини.

Нулевое распределение

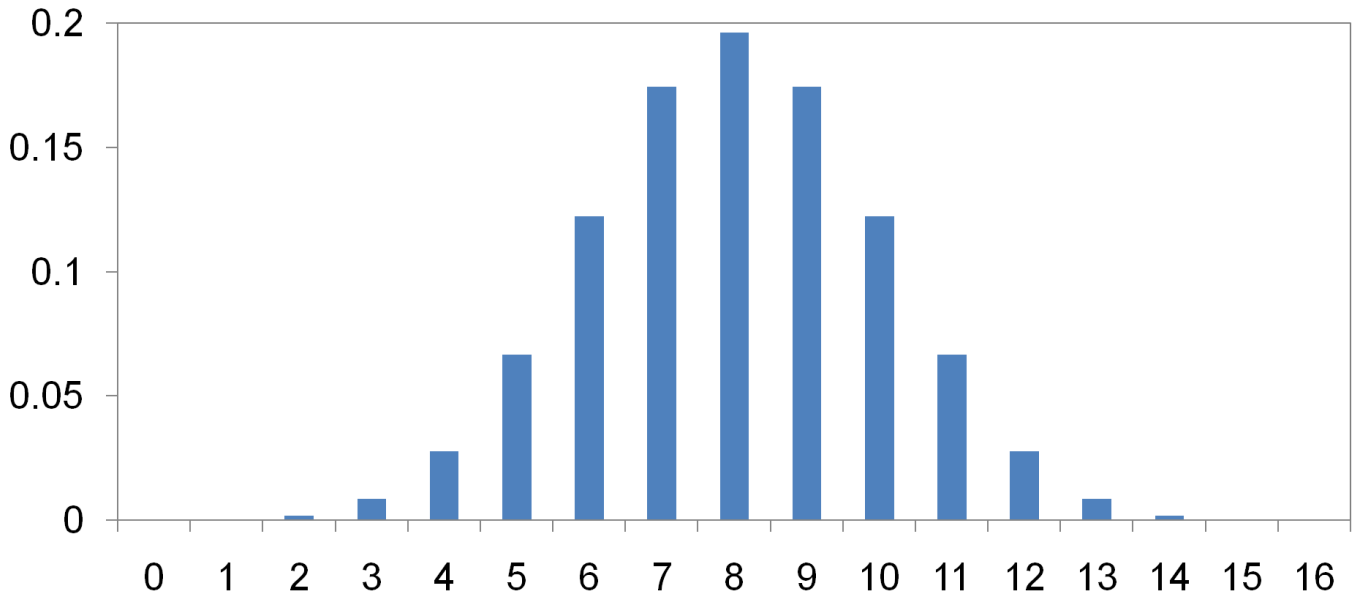
Если нулевая гипотеза справедлива и Джеймс Бонд не различает два вида мартини, то равновероятны все выборки длины n из нулей и единиц.

Пусть $n = 16$, тогда существует $2^{16} = 65536$ равновероятных варианта.

0000000000000000	0100000000000000	1000000000000000	1100000000000000
0000000000000001	0100000000000001	1000000000000001	1100000000000001
0000000000000010	0100000000000010	1000000000000010	1100000000000010
0000000000000011	0100000000000011	1000000000000011	1100000000000011
0000000000000100	0100000000000100	1000000000000100	1100000000000100
0000000000000101	0100000000000101	1000000000000101	1100000000000101
0000000000000110	0100000000000110	1000000000000110	1100000000000110
0000000000000111	0100000000000111	1000000000000111	1100000000000111
0000000000001000	0100000000001000	1000000000001000	1100000000001000
0000000000001001	0100000000001001	1000000000001001	1100000000001001
0000000000001010	0100000000001010	1000000000001010	1100000000001010
0000000000001011	0100000000001011	1000000000001011	1100000000001011
...	...	...	...

Нулевое распределение

Статистика t принимает значения от 0 до 16.

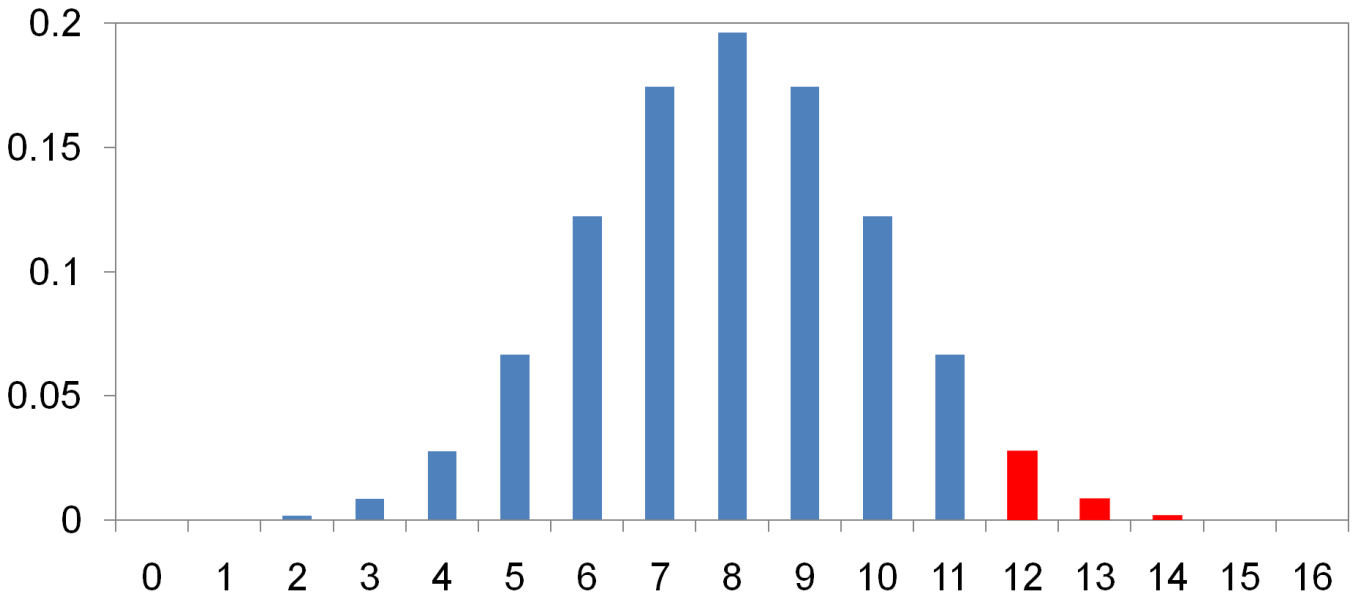


Односторонняя альтернатива

Проверим нулевую гипотезу против альтернативы H_1: Джеймс Бонд предпочитает смешанный мартини.

При справедливости такой альтернативы более вероятны большие значения статистики t (большие значения t свидетельствуют против H_0 в пользу H_1).

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



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

Достигаемый уровень значимости

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

0.0384 — вероятность реализации достаточно большого значения статистики $(t \geq 12)$ **при условии, что нулевая гипотеза справедлива**, т. е., Джеймс Бонд выбирает мартини наугад.

Достигаемый уровень значимости нельзя интерпретировать как вероятность справедливости нулевой гипотезы!

Достигаемый уровень значимости

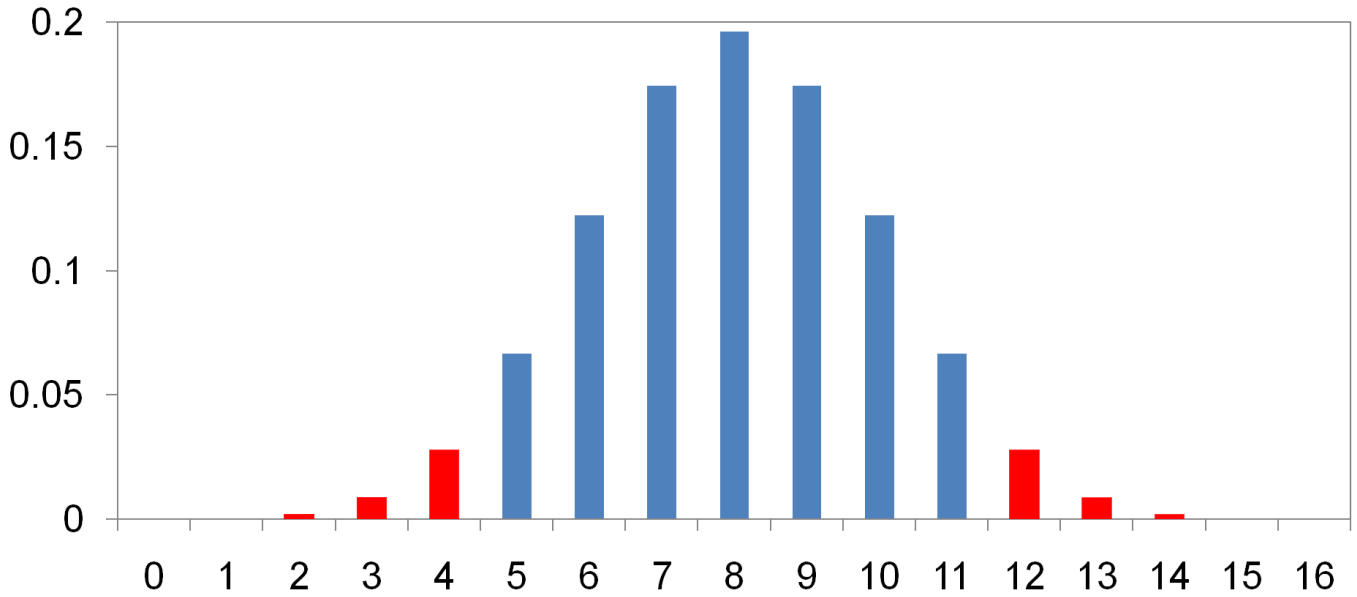
Пример: утверждается, что осьминог предсказывает результаты матчей чемпионата мира по футболу с участием сборной Германии, выбирая кормушку с флагом страны-победителя. По результатам 13 испытаний ему удаётся верно угадать результаты 11 матчей. Соответствующий достигаемый уровень значимости — $p \approx 0.0112$.

0.0112 — не вероятность того, что осьминог выбирает кормушку наугад! Исходя из здравого смысла, стоит ожидать, что эта вероятность близка к единице.

0.0112 — вероятность того, что осьминог угадает победителя в 11 и более матчах из 13 при выборе наугад.

Двусторонняя альтернатива

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



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

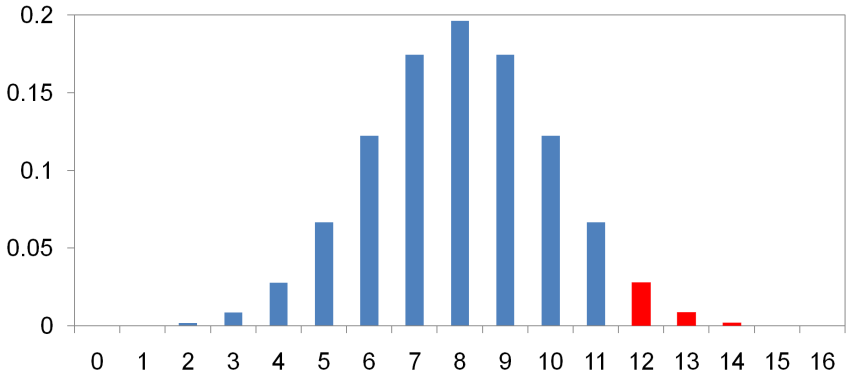
Интерпретация больших значений p-value

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

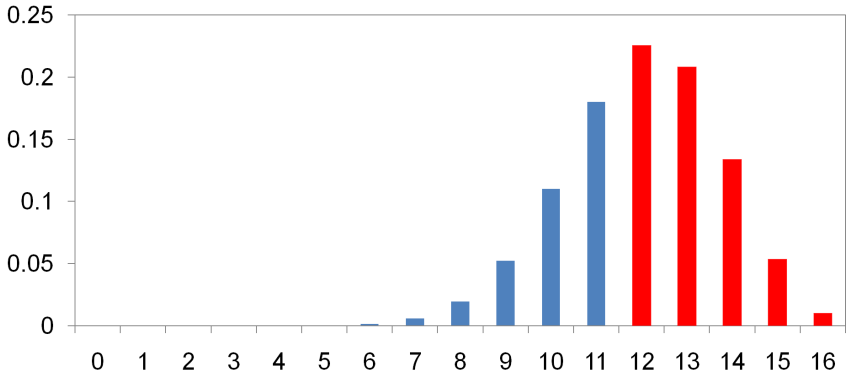
Пусть по итогам 100 испытаний Джеймс Бонд выбрал смешанный мартини 49 раз. Достигаемый уровень значимости против односторонней альтернативы — $p \approx 0.6178$. Нулевая гипотеза не отвергается, при этом сказать, что она верна, было бы ошибкой — Джеймс Бонд выбирает смешанный или взболтанный мартини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

Мощность

Проверяя нулевую гипотезу о случайном выборе мартини против односторонней альтернативы, мы отвергаем H_0 при значениях $t \geq 12$, что обеспечивает нам достигаемый уровень значимости p меньше $\alpha = 0.05$.



Пусть Джеймс Бонд выбирает смешанный мартини в 75% случаев (ненаблюдаемая вероятность).



Мощность равна ≈ 0.6302, т.е., при многократном повторении эксперимента гипотеза будет отклонена только в 63% случаев.

Мощность

Мощность зависит от следующих факторов:

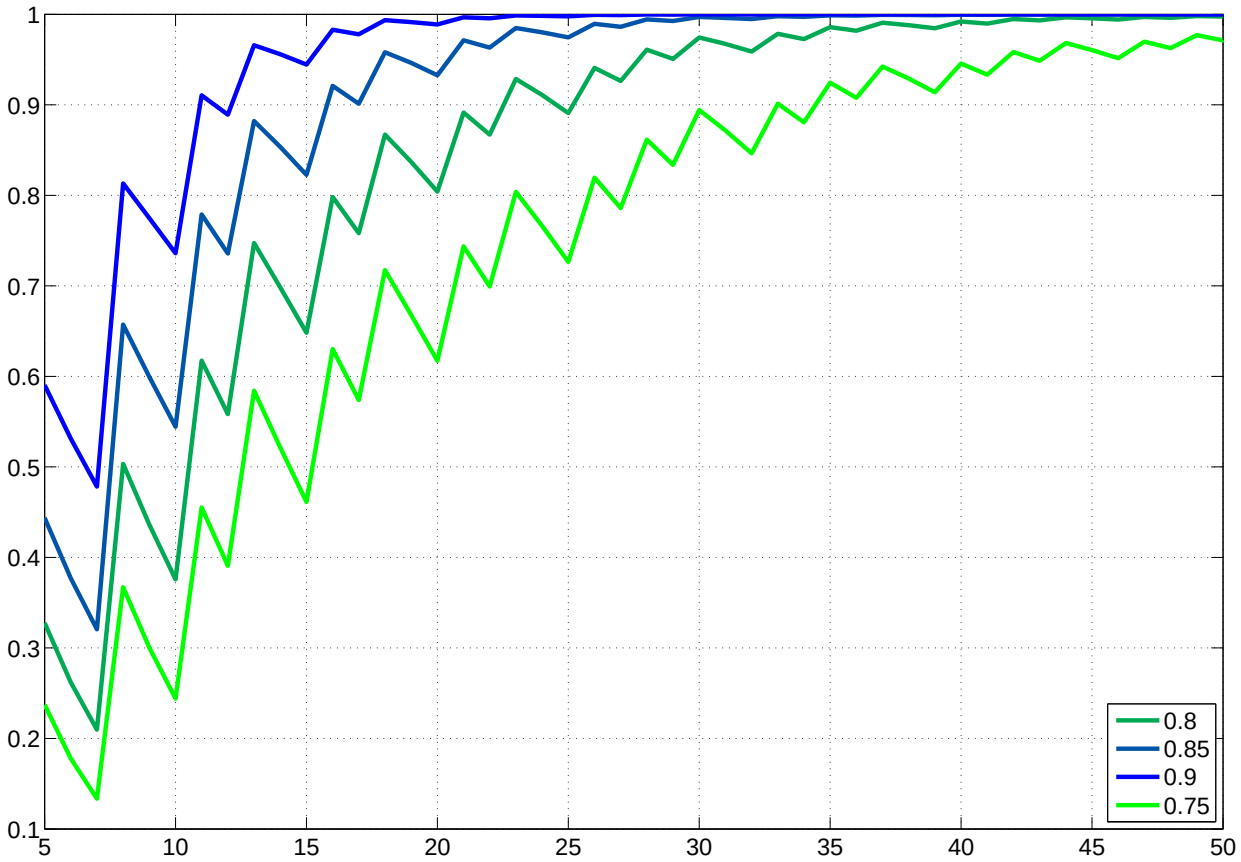
- альтернатива (односторонняя или двусторонняя)
- чувствительность статистики критерия
- размер отклонения от нулевой гипотезы
- размер выборки

Размер выборки

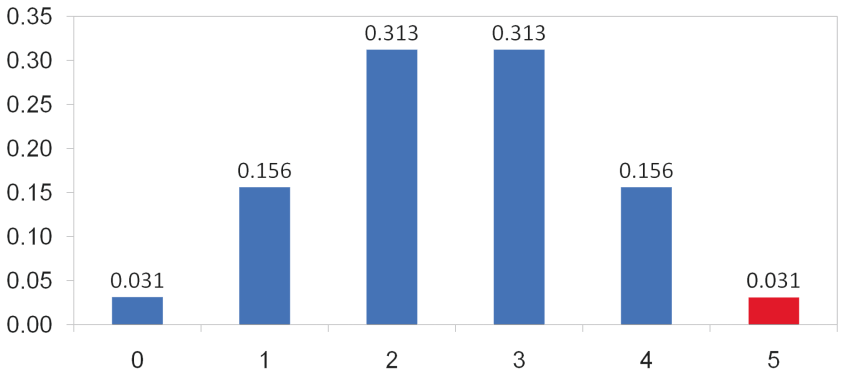
Нематематические соображения: 1 порция мартини содержит 55 мл джина и 15 мл вермута — суммарно около 25 мл спирта. Смертельная доза алкоголя при массе тела 80 кг составляет от 320 до 960 граммов спирта в зависимости от толерантности (от 13 до 38 мартини).

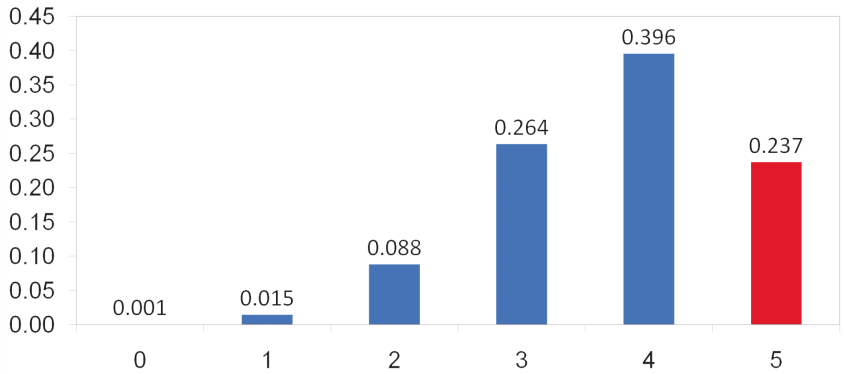
Обеспечение требуемой мощности: размер выборки подбирается так, чтобы при размере отклонения от нулевой гипотезы не меньше заданного (например, вероятность выбора смешанного мартини не меньше 0.75) мощность была не меньше заданной.

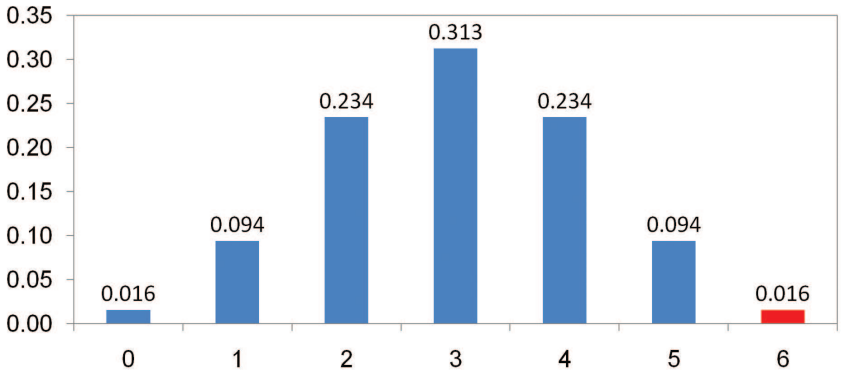
Шокирующий график

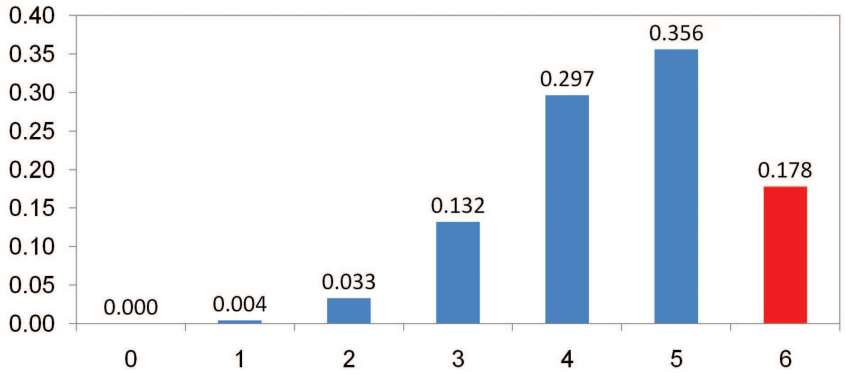


Падение мощности: объяснение









Прикладная статистика
1. Введение.

Рябенко Евгений
riabenko.e@gmail.com