Занятие 1
Погрешности, доверительные интервалы, проверка статистических гипотез

- Информация о курсе
- Виды погрешностей и правила округления
- Доверительные интервалы; нормальное распределение и t-распределение
- Проверка статистических гипотез; критерии Стьюдента (t), Фишера (F) и Пирсона (χ^2)

Сведения о курсе

Название: методы обработки результатов измерений

Преподаватель: к.х.н., с.н.с. Восков Алексей Леонидович; alvoskov@gmail.com

http://td.chem.msu.ru/study/specialcourses/

http://td.chem.msu.ru/study/generalcourses/

Темы занятий

1. Погрешности, доверительные интервалы, проверка статистических гипотез
2. Основы работы в GNU Octave (клон MATLAB)
3. Метод наименьших квадратов. Линейная и нелинейная регрессия.
4. Методы глобальной оптимизации. Метод отжига, символьная регрессия и генетические алгоритмы.

Домашние задания

1. Доверительные интервалы и проверка статистических гипотез
2. Основы работы в GNU Octave
3. Регрессионный анализ

Для получения зачёта – не менее 75% баллов по домашней работе, не менее 60% за каждое задание

Необходимое программное и аппаратное обеспечение

Программы

- MS Office 97 или выше с установленным пакетом анализа данных
- GNU Octave или MATLAB
- VirtualBox 4.3 (для работы с GNU Octave)
- Просмотровщик PDF

Возможна замена MS Office for Windows на MS Office for Mac OS X или LibreOffice (пакет анализа данных будет заменен на заготовки)

«Железо»

- X86-компьютер с 1Гб RAM или более
- 7 Гб свободного места на диске
- ОС Windows, Mac OS или Linux

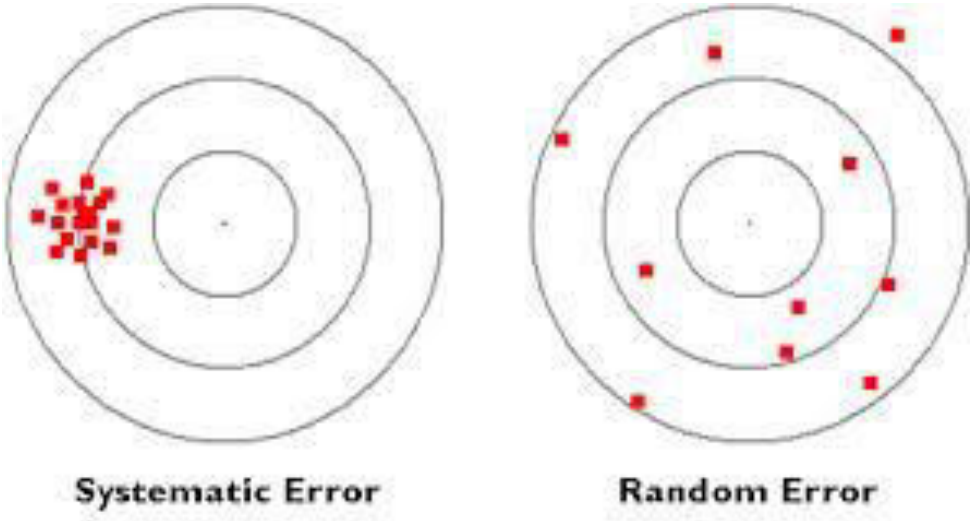
Погрешности

Виды погрешностей

- **Случайная погрешность** – вызывается большим числом причин в каждом измерении (пример – разброс между результатами титрования)
- **Систематическая погрешность** – обусловлены несовершенством метода измерений (приборы, примеси в реактивах и т.п.)
- **Грубые промахи** – связаны с ошибками экспериментатора (неправильное чтение показаний прибора и т.п.)

Абсолютная погрешность:
$\Delta x = |x_{true} - x_{meas}|$- разница между истинным и измеренным значением

Относительная погрешность:
$\delta_x = \Delta x / x$



Правила округления

Значащие цифры – все цифры данного числа от первой слева, не равной нулю, до последней справа

Примеры:

- 123 – 3 значащих цифра
- 0.012 – 2 значащих цифры
- $6.022*10^{23}$ – 4 значащих цифры
- $5*10^{3}$ – 1 значащая цифра. **НО: 5000 – 4 значащих цифры!**

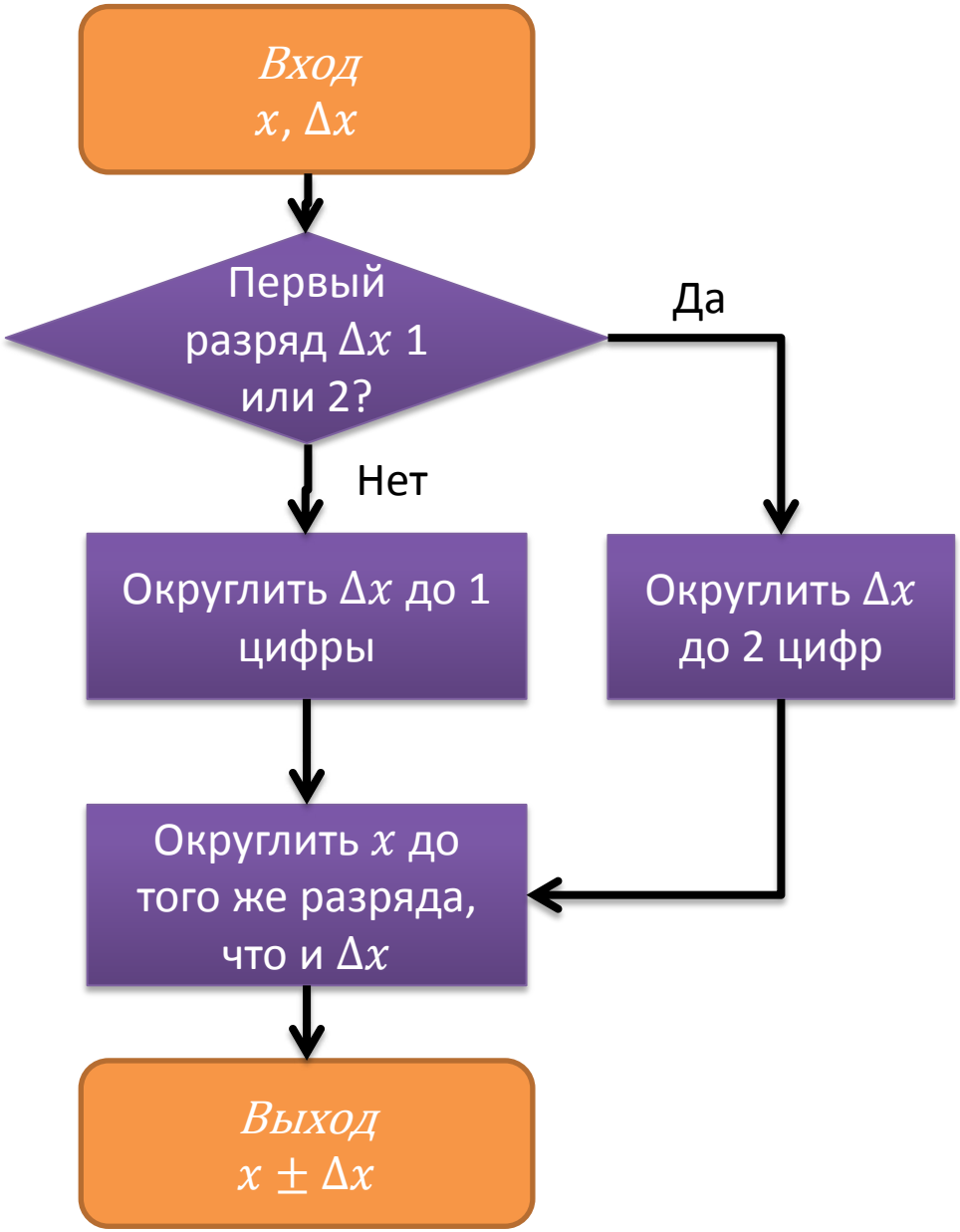
Округление до N-го разряда:

- Если N+1 – ый разряд < 5 – то отбросить все цифры после N-го разряда
- Если N+1 – ый разряд ≥ 5 – то увеличить N-ый разряд на 1 и отбросить все цифры после N-го разряда

Примеры:

- 123 -> 120
- 0.0458 -> 0.05
- 1.95 -> 2.0

Правила округления



Примеры:

- $(53216 \pm 348) \rightarrow (5.32 \pm 0.03) \cdot 10^4$
- $(0.0322 \pm 0.012) \rightarrow (3.2 \pm 1.2) \cdot 10^{-2}$
- $(12.482 \pm 0.973) \rightarrow (12.5 \pm 1.0)$

Нельзя округлять:

1. Промежуточные вычисления (потеря точности)
2. Коэффициенты регрессии, полученные МНК (они коррелированы друг с другом)

Сложение погрешностей

Сложение случайных погрешностей при сложении и вычитании:

$$\Delta y = \sqrt{\sum_i (\Delta \mathrm{x_i})^2}$$

Погрешность значения функции:

$$y = f(x_1, \dots, x_n)$$

$$\Delta y = \sqrt{\sum_i \left(\frac{\partial f(\bar{x})}{\partial x_i} \Delta x_i\right)^2}$$

Сложение систематических погрешностей при сложении и вычитании:

$$\Delta y = \sum_i |\Delta x_i|$$

Действие	Погрешность
$y = a + b;$ $y = a - b$	$\Delta y = \sqrt{(\Delta a)^2 + (\Delta b)^2}$
$y = ab;$ $y = a/b$	$\delta_y = \sqrt{\delta_a^2 + \delta_b^2}$
$y = \ln a$	$\Delta y = \delta_a$
$y = a^n$	$\delta_y = n\delta_a$
$y = \sqrt[n]{a}$	$\delta_y = \delta_a/n$

$$\delta_y = \Delta y / y$$

Доверительные интервалы

Среднее значение, стандартное отклонение, квантили

Величина	Формула	Функция MS Excel
Среднее	$\bar{x} = \frac{1}{N}\sum_i x_i$	СРЗНАЧ
Стандартное отклонение	$s_x^2 = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}}$	СТАНДОТКЛОН
Двухсторонний квантиль t-распределения	$t_{\alpha,f}$	СТЬЮДЕНТ.ОБР.2Х
Односторонний квантиль нормального распределения	z_α	НОРМ.ОБР
Левосторонний квантиль F-распределения (Фишера)	F_{α,f_1,f_2}	F.ОБР
Левосторонний квантиль хи^2-распределения (Пирсона)	$\chi^2_{\alpha,f}$	ХИ2.ОБР

Функции распределения и плотности распределения

Функция распределения вероятностей $F(x) = P(X < x)$ – вероятность того, что случайная величина X примет значение меньшее, чем *x*

Свойства:

- Определена на всей числовой прямой
- Если $x_1 < x_2$, то $F(x_1) \leq F(x_2)$
- $F(-\infty) = 0$; $F(+\infty) = 1$
- $F(x)$ непрерывна справа

Плотность распределения вероятностей непрерывной случайной величины $p(x) = \frac{dF(x)}{dx}$

Свойства:

- $\int_{-\infty}^{+\infty} p(x)dx = 1$
- $F(x) = \int_{-\infty}^{x} p(\xi)d\xi$
- $P(a < x < b) = \int_{a}^{b} p(\xi)d\xi$

Нормальное распределение

Плотность вероятности

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

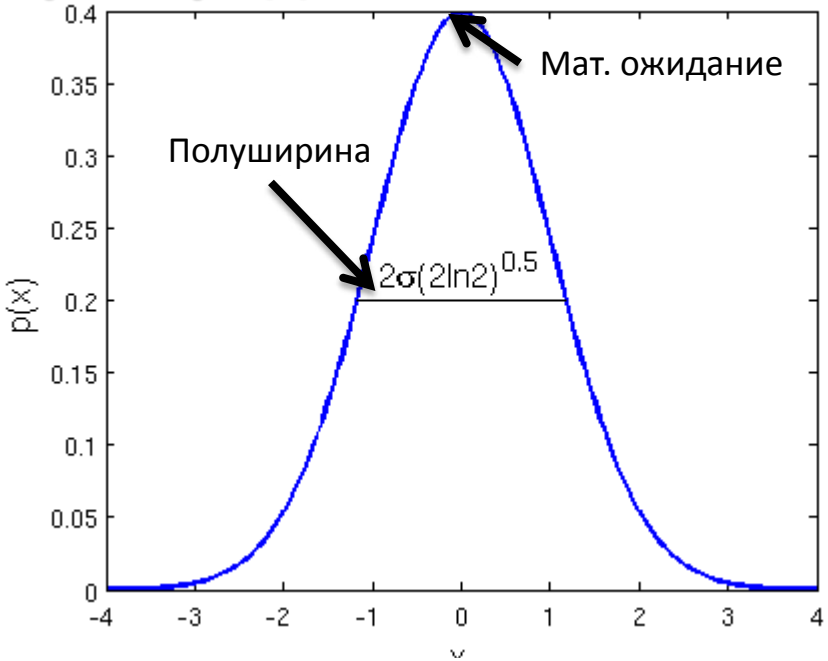
Оценка параметров нормального распределения ($n > 20$)

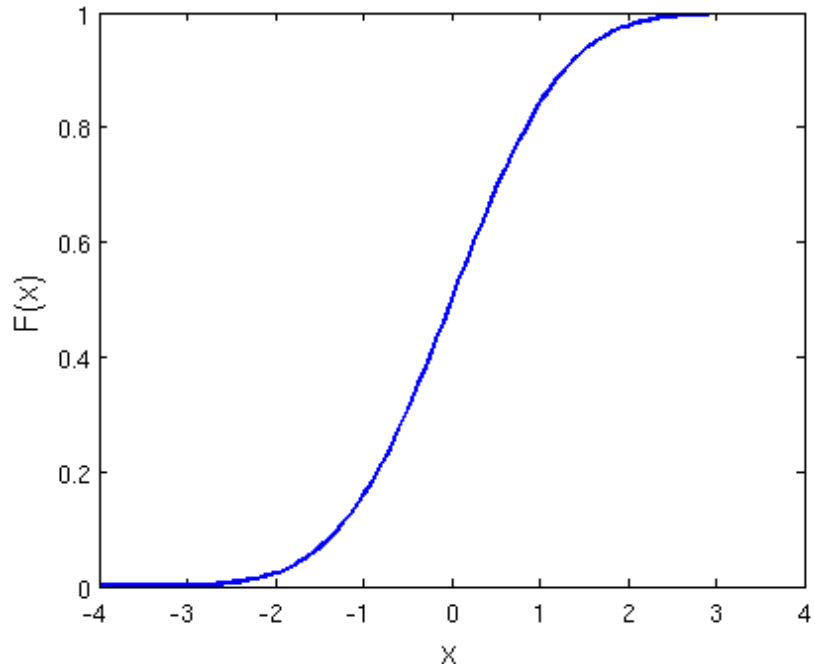
$$\mu = \bar{x} = \frac{1}{n}\sum_i x_i$$

$$\sigma = s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$

Стандартное норм. распр.

$$\sigma = 1; \mu = 0$$

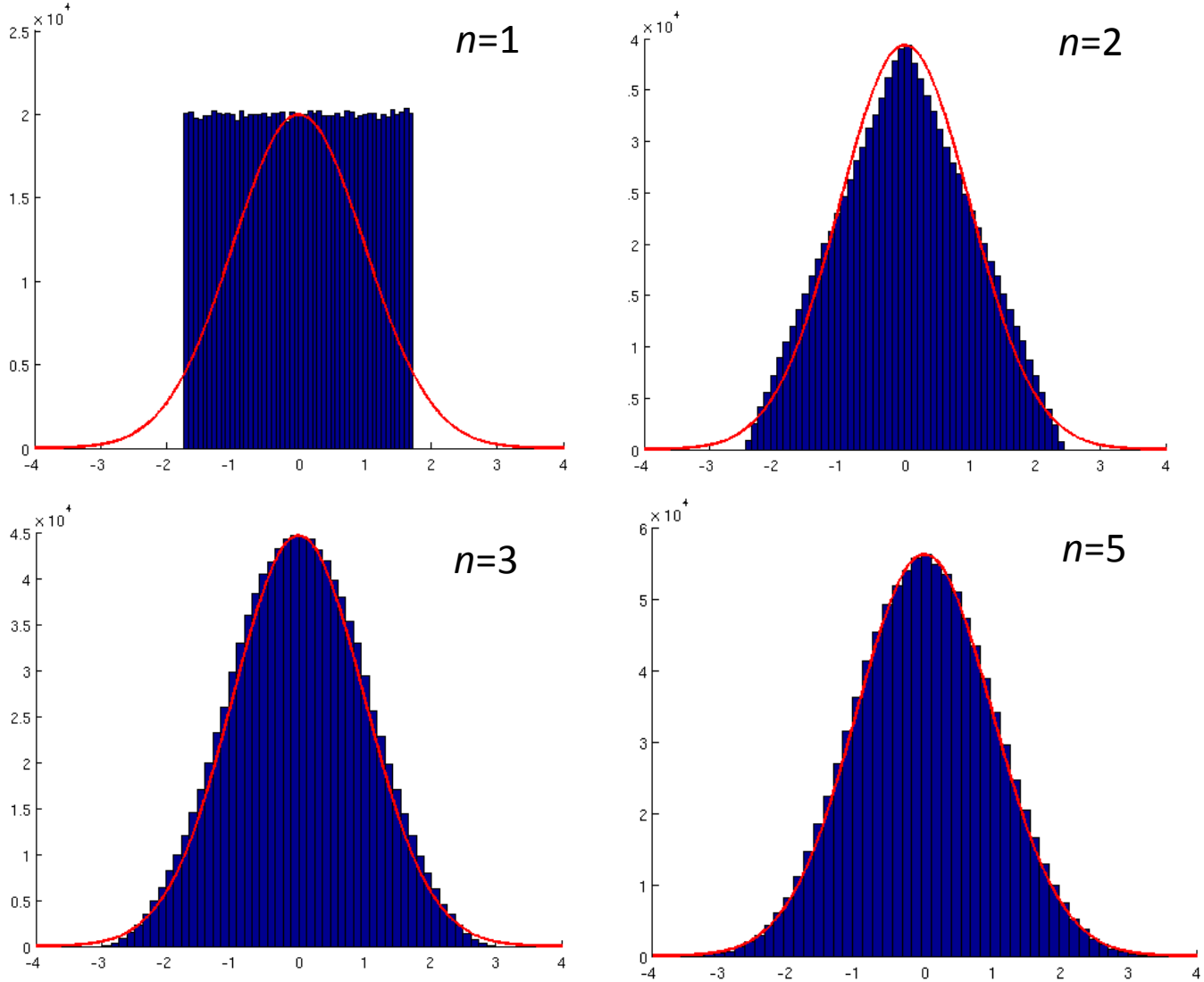




Центральная предельная теорема

Если X_i - независимые и одинаково распределенные случайные величины с конечными σ^2 и μ, то

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \to N(0;1) \quad \text{при} \quad n \to \infty$$



Распределение Стьюдента (*t*-распределение)

Плотность вероятности

$$p(y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}}$$

$$t = \frac{Y_0}{\sqrt{\frac{1}{f}\sum_{i=1}^{f} Y_i^2}}$$

Y_i – независимые стандартные нормальные случайные величины

Оценка доверительного интервала

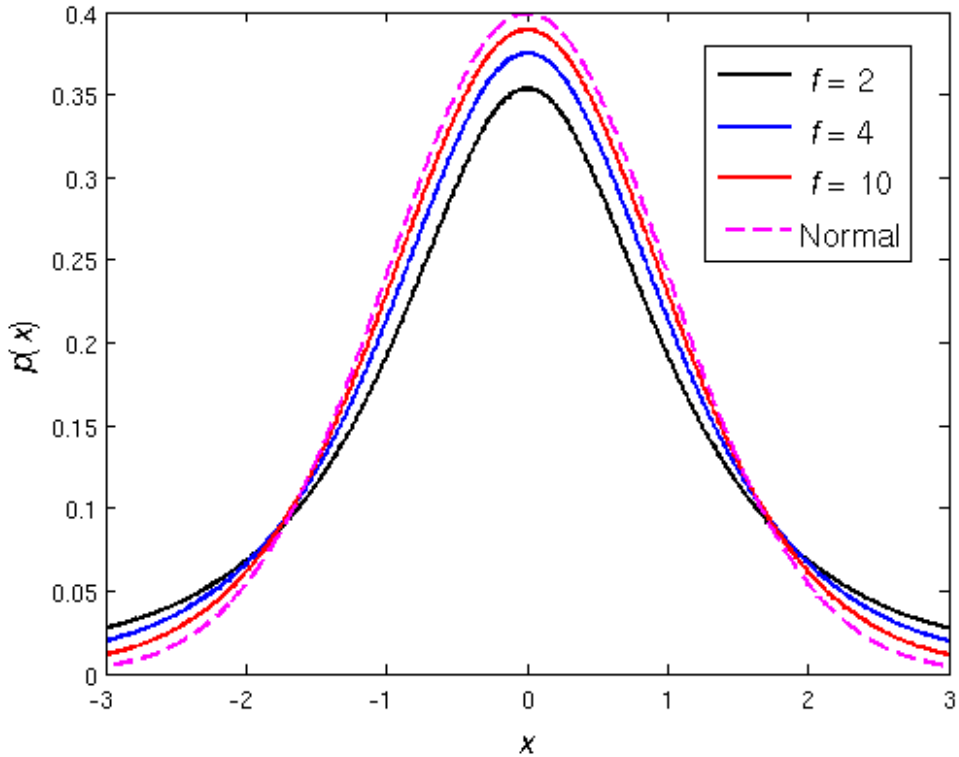
$$\bar{x} = \frac{1}{n}\sum_i x_i \quad s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$

$$t(f) = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

n – число точек

$f = n - 1$ – число степеней свободы

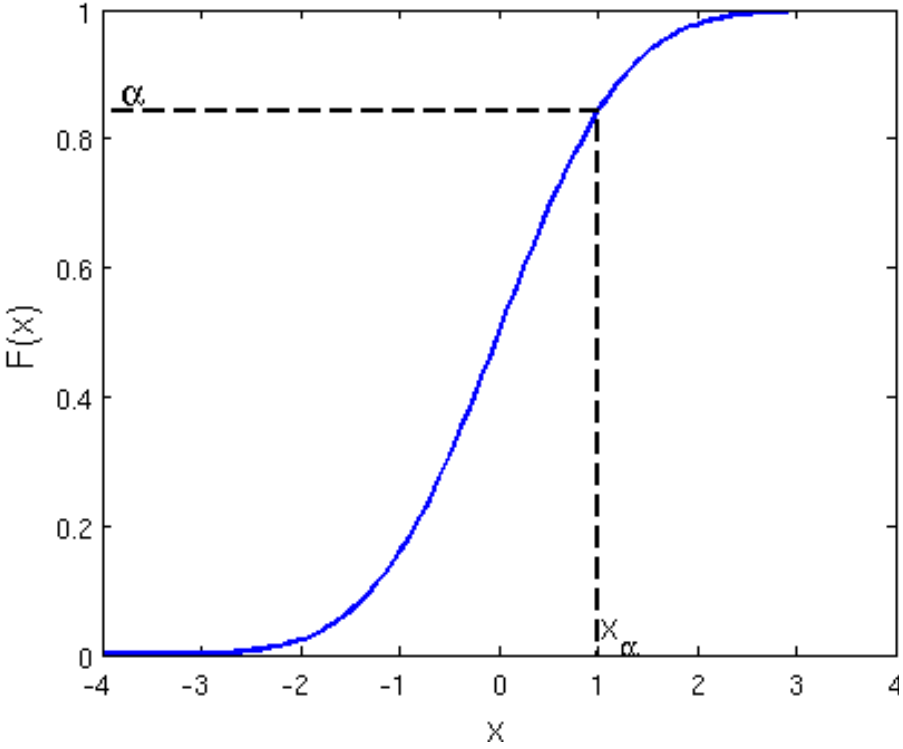
При $n \to \infty$ переходит в нормальное

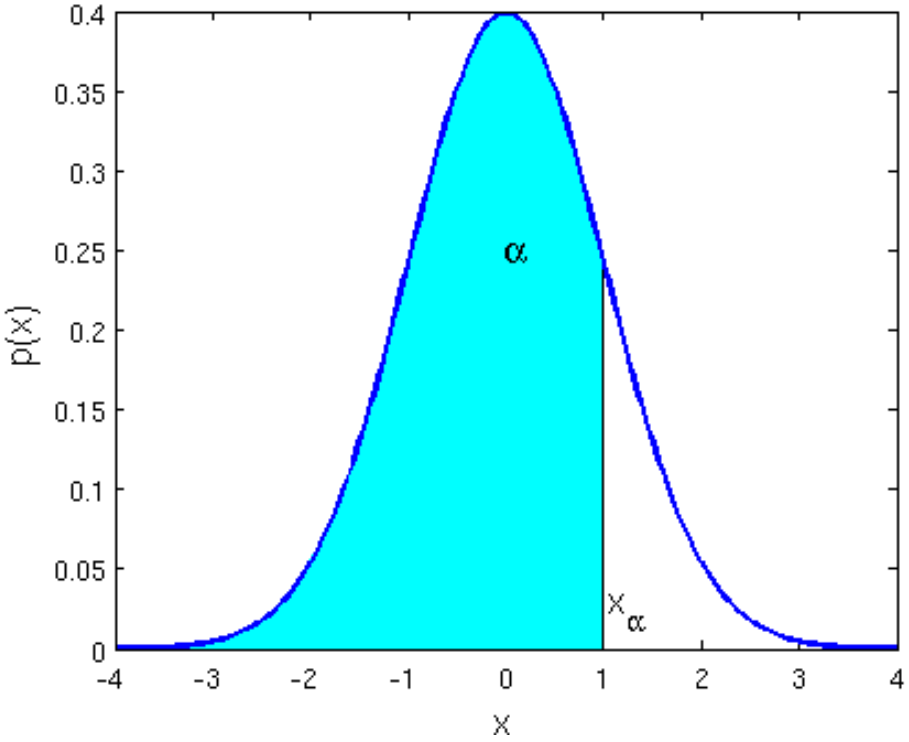


Квантили

Квантиль (α-квантиль) x_α– число, такое, что заданная случайная величина превышает его лишь с фиксированной вероятностью $(1-\alpha)$, т.е. $P(X \le x_\alpha) = \alpha$

Квантиль рассчитывается по уравнению: $F(x_\alpha) = \alpha$



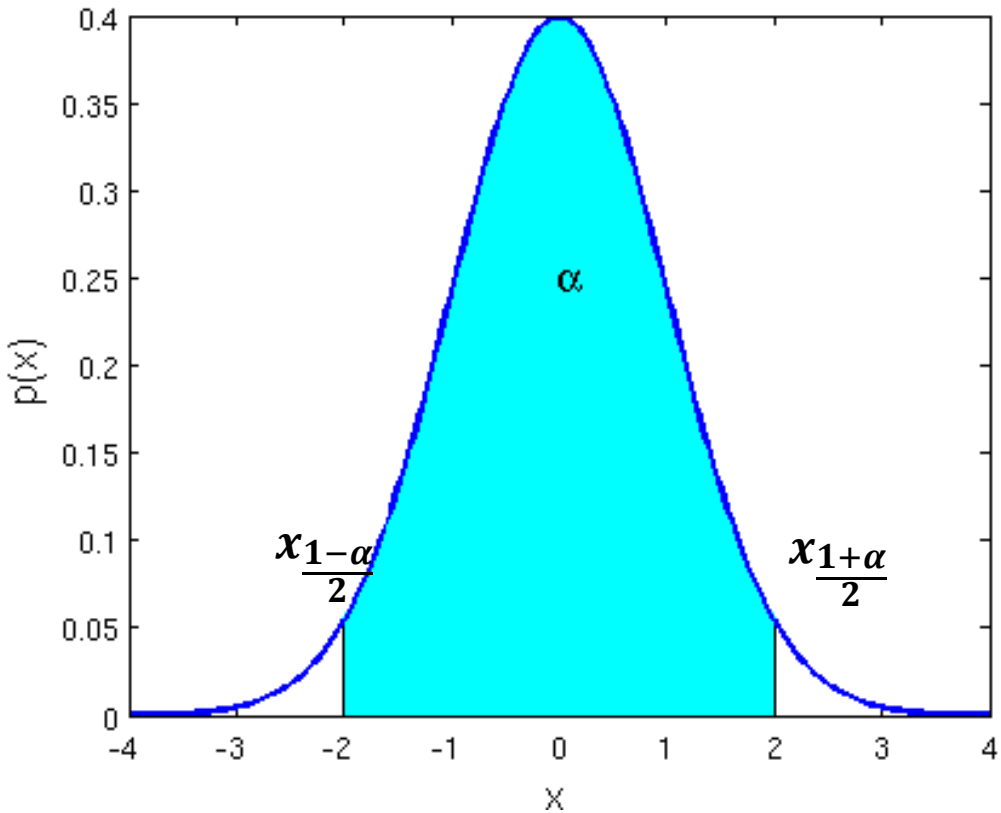


Двухсторонний квантиль

Определение

$$P\left(x_{\frac{1-\alpha}{2}} \le X \le x_{\frac{1+\alpha}{2}}\right) = \alpha$$

$$F\left(x_{\frac{1+\alpha}{2}}\right) - F\left(x_{\frac{1-\alpha}{2}}\right) = \alpha$$



Случай симметричного распределения

$$x_{\frac{1+\alpha}{2}} = -x_{\frac{1-\alpha}{2}}$$

Пример: $\alpha = 0.95$

$$\frac{1+\alpha}{2} = \frac{1+0.95}{2} = 0.975$$

$$\frac{1-\alpha}{2} = \frac{1-0.95}{2} = 0.025$$

Доверительный интервал: теория

Нормальное распределение

Если $X_1, \dots, X_n$ независимы друг от друга и $X_i \sim N(\mu_i, \sigma_i^2)$, то их линейная комбинация $Y = \sum_i c_i X_i$ подчиняется нормальному распределению $N\left(\sum_i c_i \mu_i, \sum_i c_i^2 \sigma_i^2\right)$

Распределение выборочного среднего (оценки мат.ожидания)

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

Оценка доверительного интервала

$$P\left[\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{\frac{1+\alpha}{2}} \le \mu \le \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{\frac{1+\alpha}{2}}\right] = \alpha$$

Обычно $\alpha = 0.95$ и $z = 1.96$
(«две сигмы»)

<u>ВНИМАНИЕ!</u>
Зауженный доверительный интервал при $\sigma^2 = s^2$ и $n < 50$ (особенно при $n < 8 - 10$)

При малых *n* пользуйтесь распределением Стьюдента

Доверительный интервал: теория

Распределение Стьюдента

Теорема Фишера для нормальных выборок

Если $X_1, \dots, X_n$ независимы друг от друга и $X_i \sim N(\mu, \sigma^2)$, а $\bar{X} = \frac{1}{n}\sum_i X_i$ и $s^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$, тогда

- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0;1)$ (стандартное нормальное распределение)
- $\bar{X}$ и s^2 независимы
- $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ (распределение хи-квадрат с n-1 степенями свободы)

Оценка доверительного интервала

$$P\left[\bar{X} - \frac{s}{\sqrt{n}} \cdot t_{\frac{1+\alpha}{2},f} \le \mu \le \bar{X} + \frac{s}{\sqrt{n}} \cdot t_{\frac{1+\alpha}{2},f}\right] = \alpha$$

$f = n - 1$ – число степеней свободы

Обычно $\alpha = 0.95$ и $t = 2 - 7$

<u>ВНИМАНИЕ! НЕ ПУТАТЬ!</u>

- α и $1 - \alpha$
- Одно- и двухсторонние квантили
- n и f

<u>Проверка:</u> $\lim_{f \to \infty} t_{0.95,f} = 1.96$

Доверительные интервалы: практика

1. Рассчитать $\bar{x}$ (среднее значение) и s (стандартное отклонение)

Функции MS Excel: `СРЗНАЧ`, `СТАНДОТКЛОН`

$$\mu = \bar{x} = \frac{1}{n}\sum_i x_i \qquad s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$

2. Найти двухсторонний квантиль t-распределения для заданной вероятности (обычно *p*=95%) и числа степеней свободы (*f* = *n* – 1)

Функции MS Excel: `СРЗНАЧ`, `СТАНДОТКЛОН`

(1) чем выше *p*, тем больше значение квантиля

(2) чем больше *f*, тем меньше значение квантиля

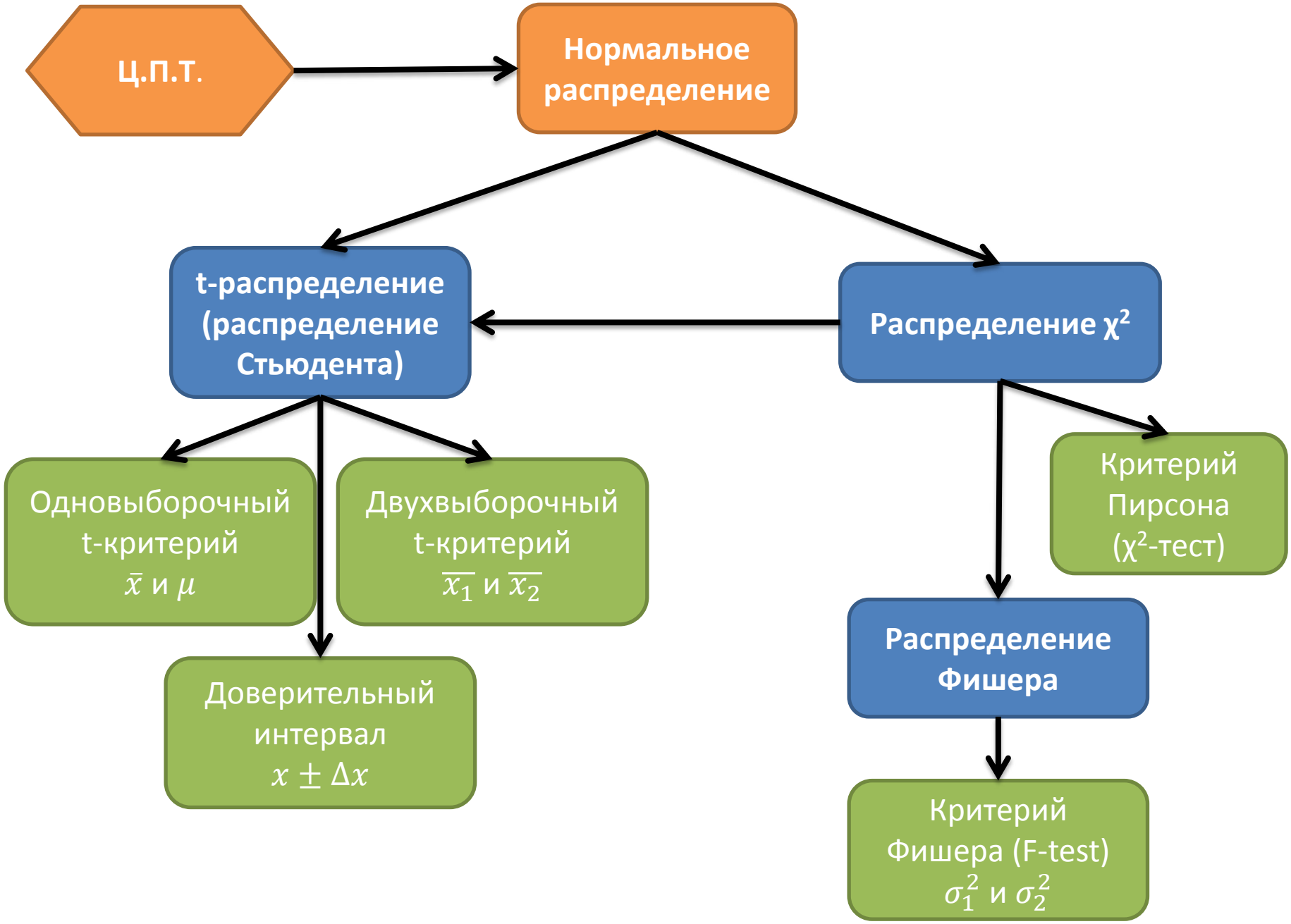
(3) для $f \approx 100$ – квантили как для нормального распределения (например, $t(p=0.95,\ f=100)=1.98$

(4) различайте *p* и 1-*p*, одно- и двухсторонние квантили!

3. Рассчитать стандартное отклонение среднего значения и доверительный интервал

$$s_{\bar{x}} = s/\sqrt{n} \qquad \Delta x = s_{\bar{x}} t(p; n-1)$$

Статистические гипотезы


Ц.П.Т.
Нормальное распределение
t-распределение (распределение Стьюдента)
Распределение χ²
Одновыборочный t-критерий $\bar{x}$ и μ
Двухвыборочный t-критерий $\overline{x_1}$ и $\overline{x_2}$
Доверительный интервал $x \pm \Delta x$
Критерий Пирсона (χ^2-тест)
Распределение Фишера
Критерий Фишера (F-test) σ_1^2 и σ_2^2

Одновыборочный t-критерий

Пусть

$\bar{x}$- среднее по выборке

μ – математическое ожидание

$s_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ - несмещённая оценка дисперсии

n – число элементов в выборке

Тогда

$$t(n-1) \sim \frac{\bar{\boldsymbol{x}} - \boldsymbol{\mu}}{s_x/\sqrt{n}}$$

Где t(n-1) – распределение Стьюдента для *n*-1 степеней свободы

<u>Дано:</u> выборка $x_1, \dots, x_n$ и математическое ожидание μ

<u>Использование критерия:</u>

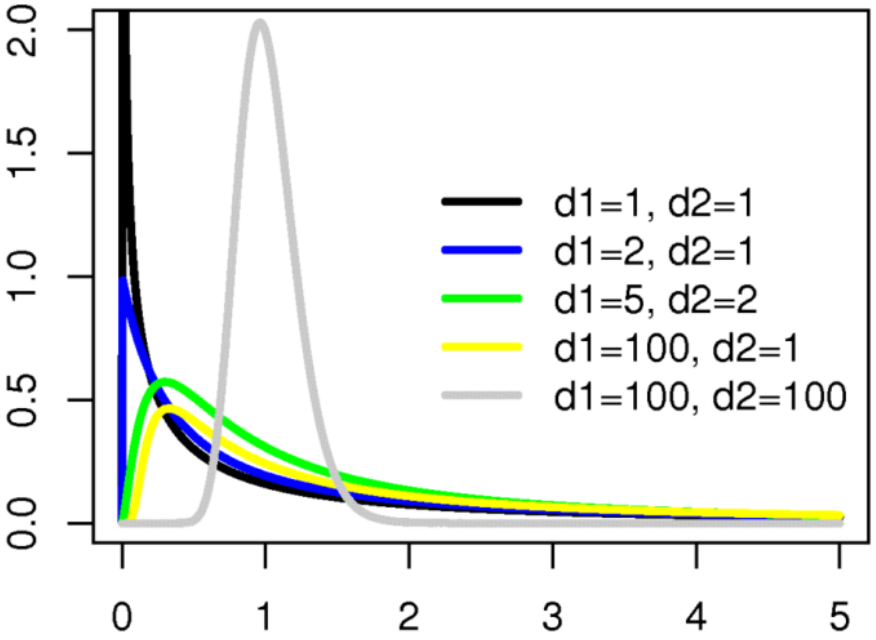
1. Рассчитать значения $\bar{x}$, s_x^2 для выборки
2. Рассчитать значение $t_{real}(n-1)=\frac{\bar{x}-\mu}{s_x/\sqrt{n}}$
3. Рассчитать t(n-1) (см. `СТЬЮДЕНТ.ОБР.2Х` в MS Excel)
4. Если $t_{real}(n-1) > t(n-1)$, то $\bar{x} \neq \mu$

<u>Примечание:</u> можно использовать функцию MS Excel `ДОВЕРИТ.СТЬЮДЕНТ`

F-распределение (Фишера)

Пусть Y_1 и Y_2 - две независимые случайные величины с распределением χ^2, т.е. $Y_i = \chi^2(d_i)$, где $d_i \in \mathbb{N}$.

Тогда $F(d_1, d_2) = \frac{Y_1/d_1}{Y_2/d_2}$ - распределение Фишера (F-распределение)



Свойства:

- Если $F \sim F(d_1, d_2)$, то $F^{-1} \sim F(d_2, d_1)$
- Если $d_1, d_2 \to \infty$, то $F \to \delta(x-1)$

Дельта-функция:

$$\delta(x) = \begin{cases} 0 \text{ если } x \neq 0 \\ +\infty \text{ если } x = 0 \end{cases}$$

F-тест (критерий Фишера)

Пусть имеются две выборки X_i $(i = 1 \dots m)$ и Y_i $(i = 1 \dots n)$ нормально распределённых случайных величин X и Y, а σ_X^2 и σ_Y^2 - выборочные дисперсии

Тогда $F = \frac{\sigma_X^2}{\sigma_Y^2} \sim F(m-1, n-1)$

1. Рассчитать стандартные отклонения s_x^2, s_y^2 для выборок X и Y

2. Если $s_x^2 < s_y^2$, то поменять выборки местами

3. Рассчитать $F_{emp} = \frac{s_x^2}{s_y^2}$ и $F(\alpha; m-1, n-1)$

Если $F_{emp} < F$, то дисперсии одинаковы

Функции MS Excel: `F.ТЕСТ`, `F.РАСП`, `F.ОБР`, `ФТЕСТ`, `ФОБР`, `FРАСП`, `FРАСПОБР`

Двухвыборочный t-критерий

$$t_{\text{эмп}}(p; df) = \frac{\bar{x} - \bar{y}}{\sigma_{x-y}}$$

Функции MS Excel: пакет анализа данных

Одинаковые дисперсии (по критерию Фишера)

$$\sigma_{x-y} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$df = n_1 + n_2 - 2$$

Разные дисперсии (по критерию Фишера)

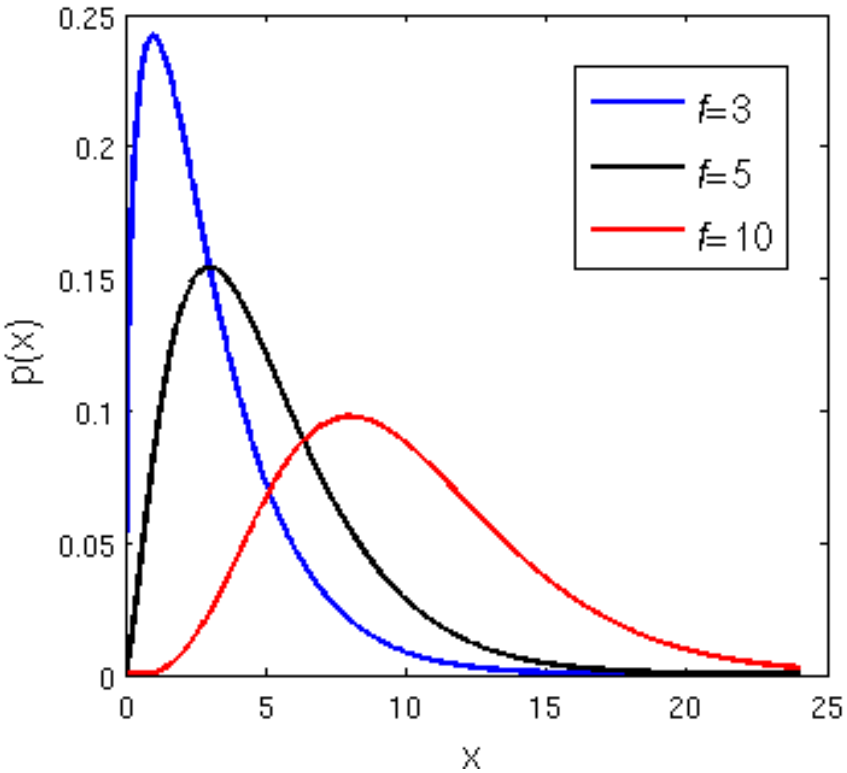
$$\sigma_{x-y} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

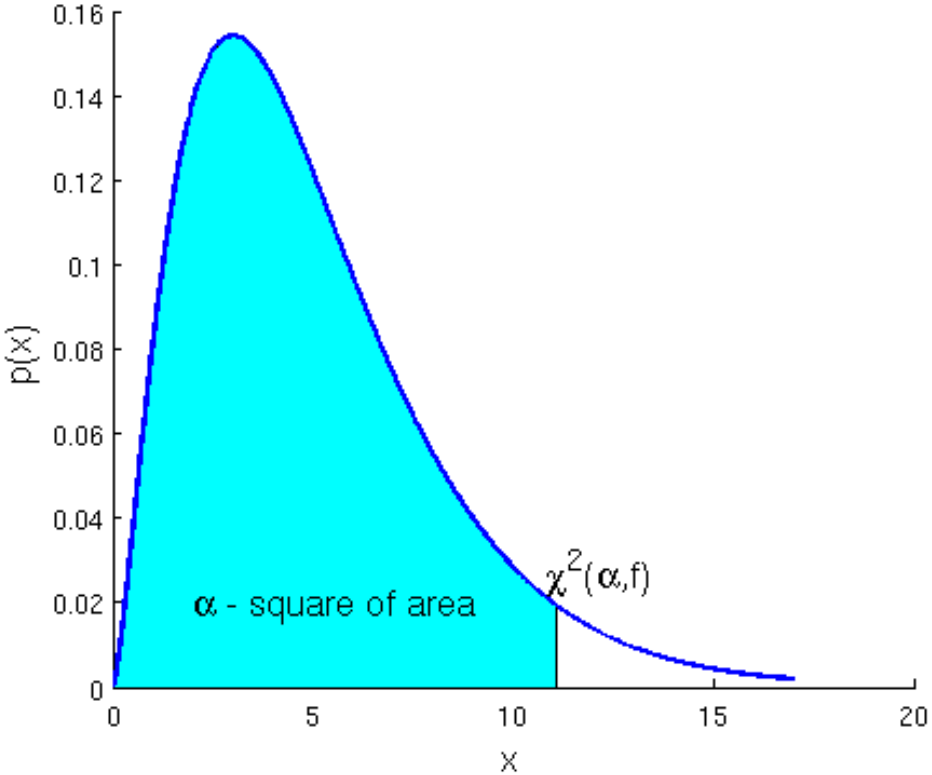
Распределение хи-квадрат (χ^2)

Пусть $z_1, \dots, z_k$ - независимые стандартные нормальные случайные величины (т.е. $z_i \sim N(0;1)$)
Тогда величина $x = \sum_i z_i^2$ имеет распределение χ^2 с k степенями свободы (т.е. $x \sim \chi^2(k)$).

Функции плотности вероятности



Квантиль $\chi^2(\alpha, f)$

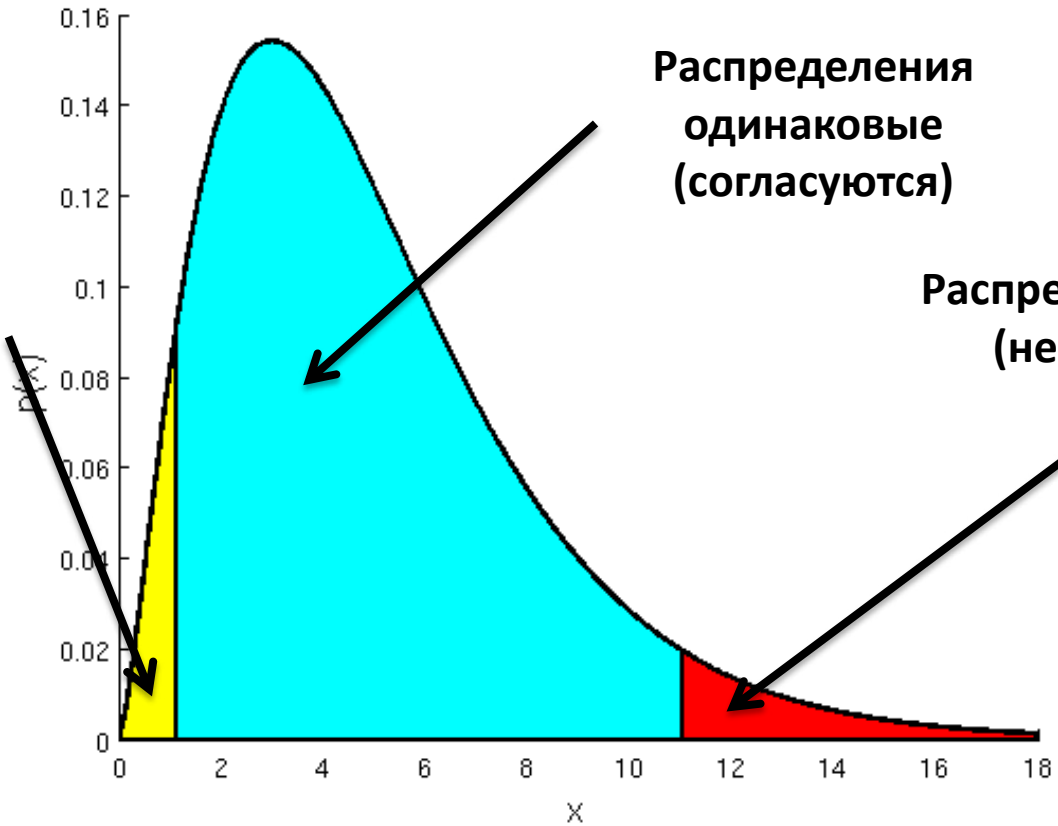


Критерий согласия χ^2 (Пирсона)

Пусть имеются 2 дискретных распределения, заданных двумя наборами частот O_i $(i = 1 \dots m)$ (наблюдаемые частоты, Observed) и E_i $(i = 1 \dots m)$ (ожидаемые частоты, Expected), причём $\sum_i O_i = \sum_i E_i$.

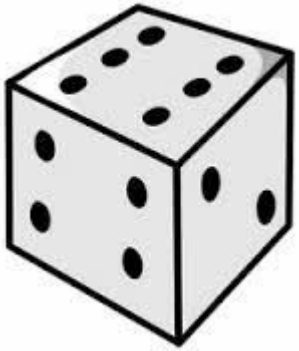
Тогда если $\chi^2_{emp} = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} < \chi^2(\alpha, m-1)$, то с вероятностью α наблюдаемое распределение совпадает с ожидаемым

«Слишком хорошее» согласие? Возможно, систематическая ошибка или подлог?



Распределения одинаковые (согласуются)

Распределения разные (не согласуются)

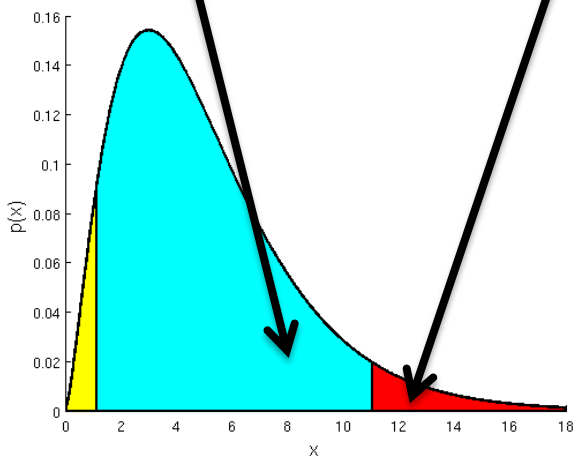


Критерий согласия χ^2 (Пирсона)
Пример с игральной костью

Игральная кость: $p_i = 1/6$

No	Oi	Ei
1	12	8
2	4	8
3	6	8
4	8	8
5	7	8
6	11	8
	48	48

```
chi2(empirical):  5.75000
chi2(a=0.95;f=5): 11.07050
chi2(a=0.05;f=5); 1.14548
```



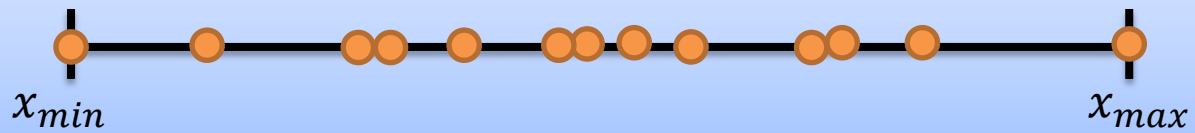
Игральная кость: $p_1 = 3p_i$ (i=2..6)

No	Oi	Ei
1	20	8
2	4	8
3	5	8
4	10	8
5	5	8
6	4	8
+	48	48

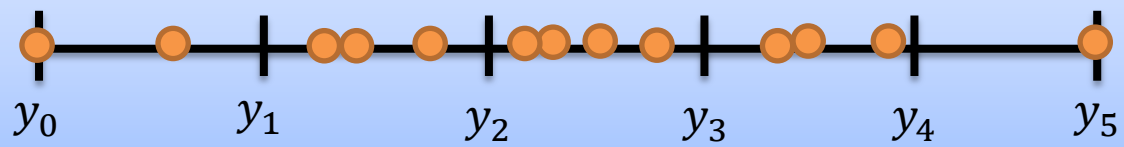
```
chi2(empirical):  24.75000
chi2(a=0.95;f=5): 11.07050
chi2(a=0.05;f=5); 1.14548
```

Критерий согласия χ^2 : непрерывное распределение

1. Найти минимальное x_{min} и максимальное x_{max} значение в выборке x_i



2. Разделить отрезок на 5-6 равных промежутков, рассчитать O_i для каждого из них (т.е. построить гистограмму)



3. Построить теоретическую гистограмму E_i (например, на основе s^2 и $\bar{x}$)

$$E_i = N \int_{y_{i-1}}^{y_i} p(x)dx = N(F(y_i) - F(y_{i-1}))$$

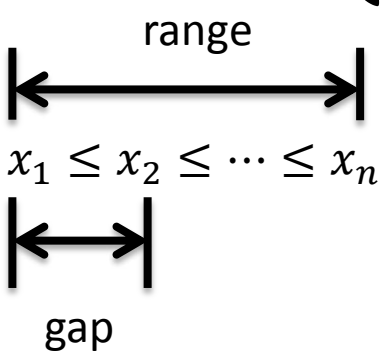
N – число точек, n – число промежутков (карманов, корзин); $y_0 = -\infty, y_n = +\infty$

4. Применение критерия Пирсона

$$\chi^2_{emp} = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Критерии для отсева грубых промахов

Грубые промахи
Q-критерий (Dixon's q-test)



$$Q = \frac{gap}{range} = \frac{|x_2 - x_1|}{|x_n - x_1|}$$

Особенности:

- Если $Q \geq Q_{tabl}$, то значение – промах
- n = 3-10
- Использовать только один раз для выборки

n	p		
	0.90	0.95	0.99
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

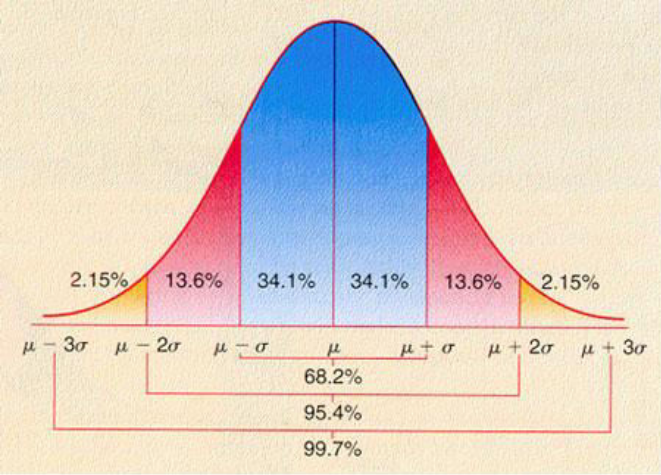
Задача: выявить промах в выборке (p=0.9):

0.189, 0.167, 0.187, 0.183, 0.186, 0.182, 0.181, 0.184, 0.181, 0.177

Грубые промахи
Критерий 3σ

Алгоритм

1. Рассчитать среднее значение
2. Рассчитать стандартное отклонение (исключив предполагаемый промах)
3. Если предполагаемый промах за пределами 3s, то исключить его
4. Применять для *n*=20-100



Задача: найти промах в выборке

8,07	8,06	8,09
8,05	8,04	8,14
8,10	8,11	8,12
8,16	8,09	8,13
8,18	8,14	8,18
8,14	8,11	8,20
8,06	8,15	8,17
8,10	8,16	
8,22	8,50	