СОВРЕМЕННЫЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Грибанов К.Г. для группы МЕНМ-190501

## Случайные величины и их распределения

Теория вероятностей и математическая статистика составляют фундамент знаний, необходимых для понимания природы любых данных, с которыми приходится иметь дело специалистам из области естественных наук.

Поскольку события реального мира не являются числами, то для того чтобы связывать события и числа, чтобы использовать математический аппарат, необходимо в общих чертах иметь представление о теории множеств. Практически, мы все всегда манипулируем понятиями теории множеств, когда приписываем событию или объекту реального мира какое-нибудь число, целое или действительное. Простой пример: алфавитный список группы студентов, посещающих лекции по данному предмету ставит в соответствие реальному человеку порядковый номер в списке, т. е. целое число. Если всех студентов взвесили, то поставили в соответствие действительное число, округленное до точности измерения, т. е. вес, например в килограммах и десятых килограмма. Если в первом случае каждому студенту поставлено в соответствие уникальное число, т. е. номер, то в случае взвешивания может оказаться, что у кого-то вес совпадет и число не будет уникальным. Здесь неявно приведено два примера разных отображений из теории множеств.

*Случайной величиной* называется величина (число), которая в результате одного наблюдения (опыта) может принимать случайным образом только одно значение, при этом условия проведения опыта фиксированы для серии наблюдений (опытов). Случайную величину в данном курсе будем обозначать, как это делается во множестве учебников, прописной латинской буквой, например *X*, а её значения (называющиеся *реализациями*), получаемые при наблюдении строчной буквой *x*. СВ может быть дискретной, непрерывной, скалярной или векторной. *Вероятность* это мера того, как часто та или иная СВ наблюдается в серии экспериментов, это тоже число, но действительное и обязательно лежащее в интервале [0; 1]. *Дискретная случайная величина* может принимать значения только из некоторого конечного или счетного множества. *Непрерывная случайная величина* принимает любые значения на некотором плотном отрезке или области, например $X \in [A;B]$ , причем отрезок может быть и бесконечным.

*Распределением* (*законом распределения*) *дискретной СВ* является совокупность вероятностей всех ее возможных значений, т. е. $P(x_\nu),\ \nu = 1,2,\ldots,N$ . Примером такого распределения является биномиальное распределение, которое описывает число *m* выпадений герба в *n* бросках, если форма монеты дает вероятность выпадения герба *q*. Кроме того, данное распределение может описывать число отказов в испытании серии новых изделий. Формула биномиального распределения следующая:

$$P(m) = \frac{n!}{m!(n-m)!} q^m (1-q)^{n-m}, \quad m=0,1,2,\ldots,n$$

На Рис.1 приведен пример биномиального распределения для *m=20* и *q=0.3. (Как создать такую картинку средствами языка сценариев R изложено в отдельной презентации «Распределения дискретных и непрерывных случайных величин и знакомство с ними в R»)*

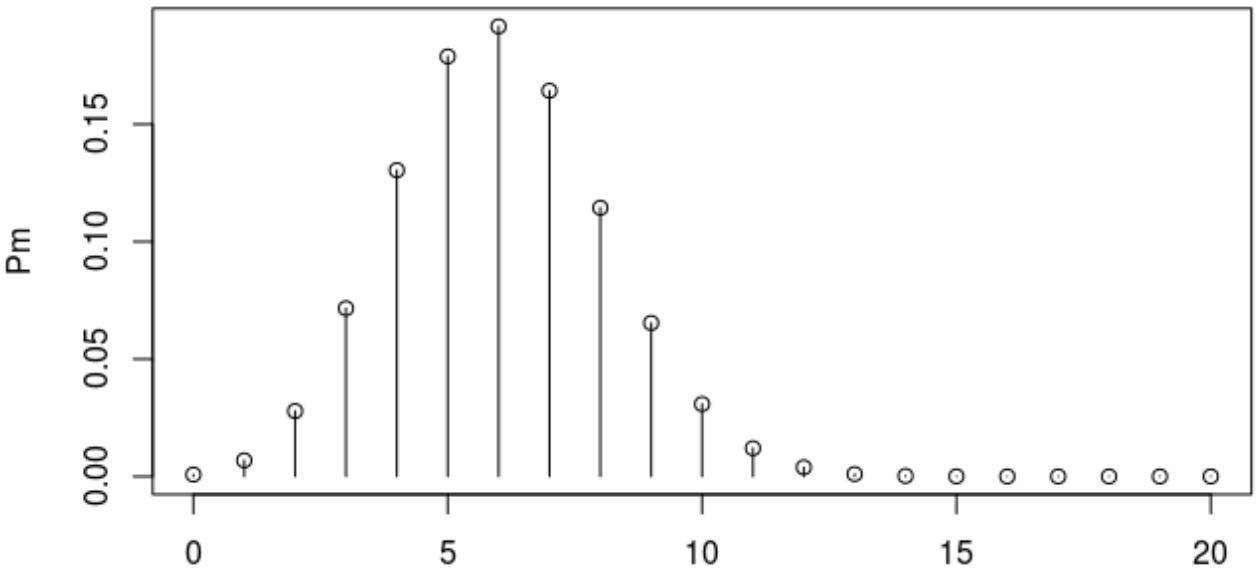


Рис. 1. Пример биномиального распределения для m=20 и q=0.3.

Вероятности дискретного распределения подчиняются условию нормировки:

$$\sum_{m=0}^{n} P(m) = 1 \ .$$

Распределение дискретной СВ можно также представить в виде ступенчатой функции распределения

$$F(m) = P\{M < m\} = \sum_{m'<m} P(m') \ ,$$

график которой приведен для того же биномиального распределения на Рис.2. Следует понимать условность такой функции и ее ступенчатого графика для дискретного распределения, поскольку значения аргумента, откладываемые по оси абсцисс на Рис.2 могут принимать только дискретные значения, а значений абсциссы между ними, на «полках» ступенчатой функции нет во множестве случайной величины.

Распределение (закон распределения) непрерывной СВ характеризуют *плотностью распределения вероятности* $f(x)$ или *функцией распределения* $F(x)$ . Плотность [распределения] вероятности скалярной СВ, является производной от функции распределения, т. е.

$$f(x) = \left(F(x)\right)' \quad .$$

Плотность вероятности позволяет вычислить вероятность попадания СВ в заданный интервал, т. е.

$$P\{X \in [x-\delta; x+\delta]\} \;=\; \int_{x-\delta}^{x+\delta} f(x')dx' \quad .$$

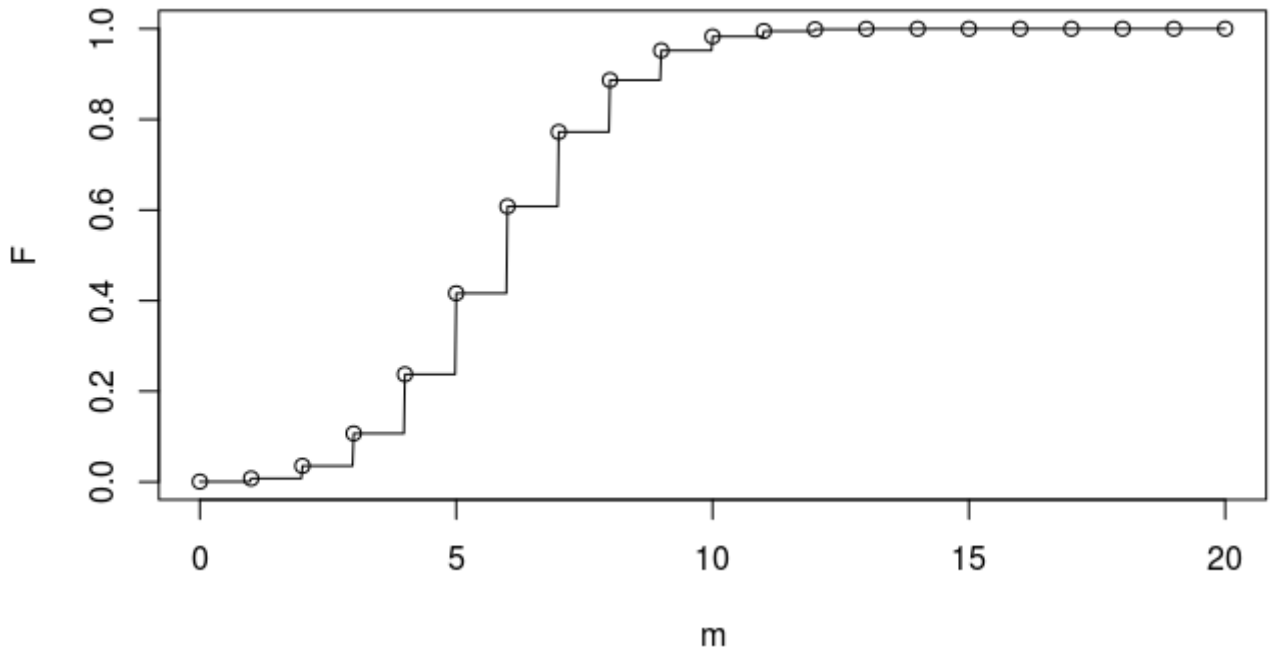


Рис. 2. Ступенчатая функция распределения для биномиального распределения, показанного на Рис. 2

Функция же распределения представляет собой вероятность того, что СВ $X \in [A; B]$ не превысит некоторое значение *x*, т.е.

$$F(x) \;=\; P\{X \le x\} \;=\; \int_{A}^{x} f(x')dx' \quad ,$$

здесь штрихованная переменная под интегралом используется для того, чтобы отличать ее от верхнего предела интегрирования. Вот некоторые свойства функций плотности вероятности и распределения:

1. $f(x) \;\ge\; 0, \forall\, x \in [A; B]$ , т. е. $f(x)$ неотрицательная;

2. $\int_A^B f(x)dx = 1$ , условие нормировки;

3. $F(A) = 0$, $F(B) = 1$;

4. Если $x_1 \geq x_2$ , то $F(x_1) \geq F(x_2)$ , т. е. функция $F(x)$ монотонно возрастает.

На Рис. 3 в качестве примера приведены $f(x)$ и $F(x)$ для гамма-распределения (с которым подробнее следует познакомиться с помощью презентации «*Распределения дискретных и непрерывных случайных величин и знакомство с ними в R*»).

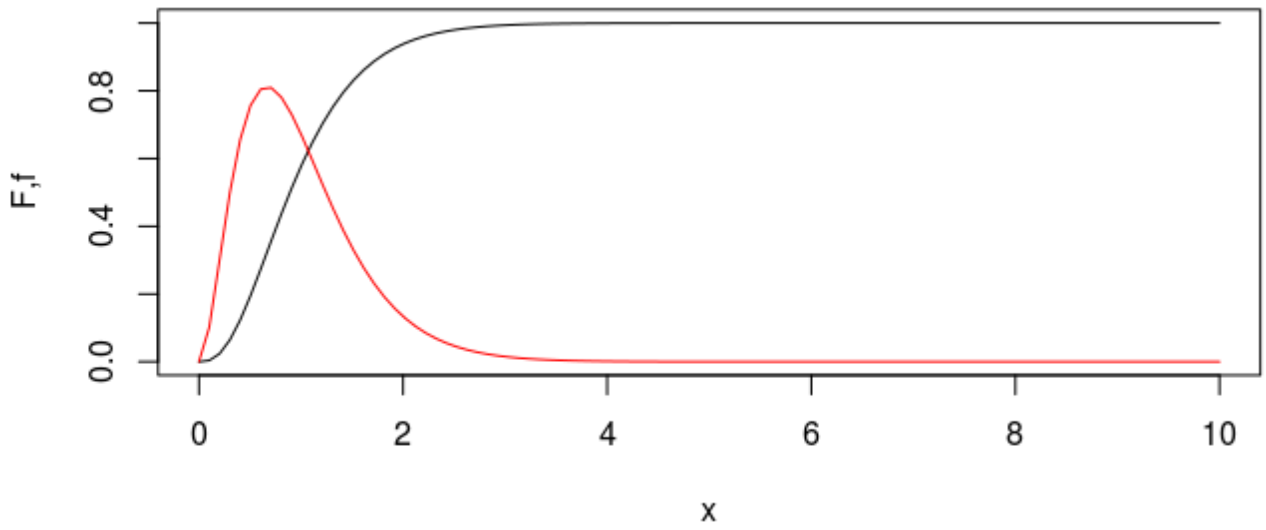


Рис. 3. Пример плотности вероятности для гамма-распределения (красная кривая) и функции распределения (черная кривая).

Векторная случайная величина это совокупность взаимосвязанных скалярных СВ, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ (здесь Т — это знак транспонирования, чтобы представлять вектор в виде матрицы столбца).

## Моменты распределения случайных величин

Для решения практических задач часто достаточно только двух первых моментов характеризующих положение — *математическое ожидание* и рассеяние — *дисперсия*. *Математическое ожидание* для дискретной СВ

$$M_X = \sum_{n=1}^{N} x_n P(x_n) \ ,$$

для непрерывной СВ

$$M_X = \int_A^B x f(x) dx \ .$$

*Дисперсия* СВ представляет собой второй центральный момент распределения и определяется следующим образом:

для дискретной СВ $$D_X = \sum_{n=1}^{N} (x_n - M_X)^2 P(x_n) \ ,$$

для непрерывной СВ $$D_X = \int_A^B (x - M_X)^2 f(x) dx \ .$$

Размерность дисперсии не совпадает с размерностью СВ, поэтому на практике используется *среднеквадратическое отклонение* (СКО) СВ:

$$\sigma_X = \sqrt{D_X} \ .$$

С учетом этой формулы, дисперсию часто обозначают как $\sigma_X^2$ . На практике часто используют «правило трех сигм», имея в виду, что все реализации СВ почти наверное попадают в интервал $[M_X - 3\sigma_X ; M_X + 3\sigma_X]$ .

Распределение векторной СВ $\boldsymbol{X} = (X_1, X_2, ..., X_n)^T$ характеризуется векторным МО и ковариационной матрицей $\boldsymbol{K_X}$ . *Математическое ожидание* векторной СВ представляет собой вектор состоящий из математических ожиданий компонентов вектора, т.е.

$$\boldsymbol{M_X} = (M_{X1}, M_{X2}, ..., M_{Xn})^T \ .$$

*Ковариационная матрица векторной СВ* состоит из дисперсий компонентов на главной диагонали и *моментов ковариации* пар компонентов в качестве внедиагональных компонентов:

$$\boldsymbol{K_X} = \begin{Bmatrix} D_{X1} & K_{X1,X2} & ... & K_{X1,Xn} \\ K_{X2,X1} & D_{X2} & ... & K_{X2,Xn} \\ ... & ... & ... & ... \\ K_{Xn,X1} & K_{Xn,X2} & ... & D_{Xn,Xn} \end{Bmatrix} \ .$$

Коэффициенты ковариации можно определить в терминах математического ожидания от произведения центрированных пар компонентов векторной СВ, т.е.

$$K_{Xi,Xj} = E\left[(x_i - M_{Xi})(x_j - M_{Xj})\right] \ ,$$

где E это оператор математического ожидания, если $i=j$ , то формула переходит в формулу для дисперсии на главной диагонали.

**Простой пример.** *Равномерное распределение* на отрезке $[A;B]$ имеет плотность вероятности

$$f(x) = \begin{cases} \frac{1}{B-A}, & x \in [A;\ B] \\ 0, & x \notin [A;\ B] \end{cases},$$

и функцию распределения

$$F(x) = \begin{cases} 0, & x < A \\ \frac{x-A}{B-A}, & x \in [A;\ B] \\ 0, & x > B \end{cases}.$$

На Рис. 4 показаны обе функции для A=1,B=6.

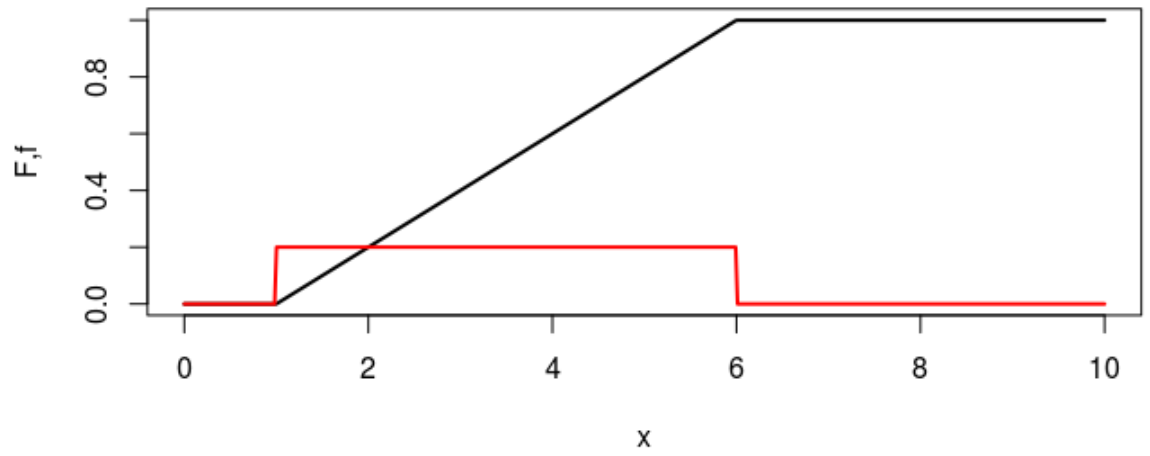


Рис. 4. Функция распределения (черная) и плотность вероятности (красная) для равномерного распределения на отрезке $[1;6]$ .

*Предлагается самостоятельно получить выражения для* $M_X$ *и* $D_X$ *для равномерного распределения на отрезке* $[A;\ B]$ *, а также воспроизвести Рис.4 средствами R.*

## Задачи математической статистики

Математическая статистика связана с теорией вероятностей и в ней ставятся и решаются задачи изучения закономерностей и получения выводов о СВ по выборке данных наблюдений, принадлежащих совокупности СВ, называемой *генеральной*.

*Генеральная совокупность* — совокупность объема *N* всех физических объектов одного вида или значений СВ с определенным распределением. Число N может быть конечно, а может и $N \to \infty$ .

*Выборка* — совокупность элементов размера $n < N$ , случайно полученная из ГС при некоторых условиях эксперимента. Выборка может быть *повторной* — отобранные элементы возвращаются в ГС, или *бесповторной* — элементы не возвращаются в ГС. Задача математической статистики по выборке получить сведения о ГС. ГС может быть *стабильной* или *динамической*, т. е. изменяющейся со временем. Чтобы не углубляться в теорию случайных процессов, в данном курсе рассматриваются только стабильные ГС. Выборка должна быть *репрезентативной* (*представительной*), что реализуется если она реализуется случайно, т. е. все элементы ГС имеют одинаковые шансы попасть в выборку. Выборка должна быть *достаточной*, т. е. иметь достаточный объем, чтобы обеспечить хорошую точность сведений о ГС. Это не всегда удается достичь и приходиться довольствоваться выборкой малого объема.

Из выборки можно получить *статистическое распределение* СВ. Для дискретной СВ $x_q, q = 1,2,3,\ldots,Q$ при наличии выборки объема *n*, статистическое распределение определяется как

$$p_q = \frac{n_q}{n}, \quad q = 1,2,\ldots,Q \ ,$$

где $n_q$ это число наблюдений величины $x_q$ в выборке объема *n*. Естественно выполнение следующих соотношений:

$$\sum_n n_q = n, \qquad \sum_n p_q = 1.$$

Статистическое распределение дискретной СВ можно отобразить на графике в стиле Рис.1.

Для непрерывной СВ, при наличии выборки $x_i, i = 1,2,3,\ldots,n$ можно построить *гистограмму распределения* или эмпирическую *функцию распределения*. Для этого диапазон реализации $\Delta = x_{max} - x_{min} + \varepsilon$ делится на K отрезков одинаковой длины $h = \Delta / K$ , K рекомендуется выбирать в интервале от 6 до 20, а в случае большой выборки $K \approx n^{1/3}$ . Значение $\varepsilon < h$ выбирается из соображений удобства представления гистограммы. Выбираются середины отрезков $x_k$, $k = 1,2,\ldots,K$ , вычисляются $n_k$ - числа попадания СВ из выборки в интервал $[x_k - h/2;\ x_k + h/2]$ , а затем вероятности попадания в этот интервал

$$P_k = \frac{n_k}{n}, \quad k = 1,2,3,\ldots,K \ .$$

Можно построить гистограмму из прямоугольников высотой $P_k$ и шириной *h*, однако если нормировать высоту прямоугольников $f_k = P_k / h$ , можно

построить статистическую плотность вероятности, площадь под гистограммой в этом случае будет удовлетворять условию нормировки

$$\sum_{k=1}^{K} f_k h = 1 \quad .$$

Для построения *эмпирической функции распределения* на основе выборки строится вариационный ряд, т. е. реализации $x_i$ записываются в возрастающем порядке, при этом получается новая последовательность $x_q, q=1,2,\ldots,n$ , которая используется для вычисления ФР:

$$F(x_q) = P\{x < x_q\} = \frac{(q-1)}{n}, \qquad q=1,2,\ldots,n \quad .$$

При $n \to \infty$, $f_k \to f(x_k)$, $F(x_q) \to F(x)$ . Примеры обоих эмпирических функций для нормального распределения показаны на Рис.5.

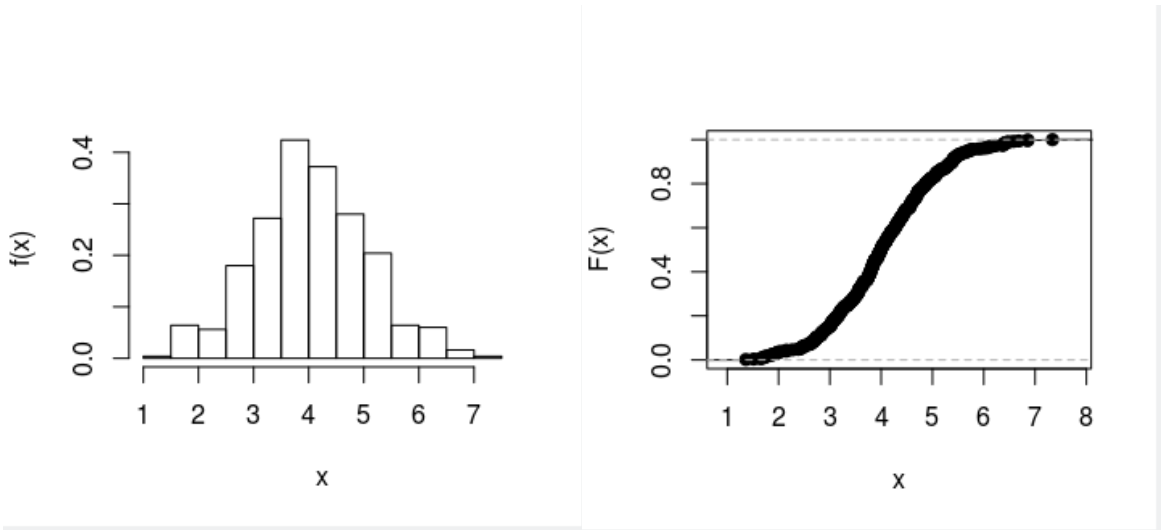


Рис.5. Примеры статистической плотности вероятности и функции распределения. Генерация выборки нормально распределенной СВ выполнена с помощью функции `rnorm()`, а построение графиков выполнено в R с помощью функций `hist()` и `ecdf()`.

## Точечное оценивание параметров распределения случайных величин

Оцениваемые на основе ограниченной выборки параметры распределения случайной величины являются неточными и называются *статистическими оценками*, которые зависят от выборки и сами являются случайными величинами. *Точечная оценка* это величина, зависящая от выборки и выражаемая одним числом или вектором. Пусть СВ имеет распределение с параметром $\theta$ , *точечную оценку* или просто *оценку* этого параметра будем

обозначать $\hat{\theta}$ . Для получение точечных оценок используются три основных метода: метод максимального правдоподобия (ММП), метод наименьших квадратов (МНК) и метод моментов (ММ). Поскольку оценка это СВ, то и она характеризуется математическим ожиданием

$$M_{\hat{\theta}} = E\left[\hat{\theta}\right]$$

и дисперсией $\sigma^2_{\hat{\theta}}$ . Иногда удается построить плотность вероятности оценок параметра $f(\theta)$ и если такая функция построена корректно, то можно определить *несмещенную точечную оценку параметра* $\theta$ и её *дисперсию* как

$$\hat{\theta} = \int_{-\infty}^{\infty} \theta f(\theta)\, d\theta \ ,$$

$$\sigma^2_{\hat{\theta}} = \int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 f(\theta)\, d\theta \ .$$

К точечной оценке предъявляются требования по её *несмещённости*, *эффективности*, *состоятельности* и *достаточности*.

*Несмещенность* оценки означает что математическое ожидание оценки равно точному параметру, т.е.

$$E\left[\hat{\theta}\right] = \theta \ .$$

*Эффективность* оценки означает, что ее дисперсия является минимальной в классе всех несмещенных оценок параметра $\theta$ :

$$\sigma^2_{\hat{\theta}} = min \ .$$

*Состоятельная* оценка сходится по вероятности к истинному значению при увеличении объема выборки, т.е.

$$P\left\{|\hat{\theta} - \theta| < \delta\right\} \xrightarrow[n \to \infty,\, \delta \to 0]{} 1 \ .$$

*Достаточной* называется оценка, которая содержит такое же количество информации о неизвестном параметре $\theta$ , что и элементы выборки.

Получение информации с помощью $f(\theta)$ затруднено тем, что такую функцию редко удается построить. Свойства оценок скалярного параметра, рассмотренные выше, можно распространить и на векторный параметр $\boldsymbol{\theta}$ .

## Интервальное оценивание

*Интервальная оценка* это статистическая оценка в виде совокупности точек или множества. Обычно это *доверительный интервал* $[\theta_H; \theta_B]$ , такой что

$$P\{\theta_H \le \theta \le \theta_B\} = \gamma ,$$

где $\gamma$ это *доверительная вероятность*, $\theta_H$ и $\theta_B$ - *нижняя* и *верхняя доверительные границы* соответственно. Нижняя или верхняя границы могут оказаться и бесконечными. Точность интервальной оценки характеризуется размахом доверительного интервала, а ее достоверность или надежность — доверительной вероятностью.

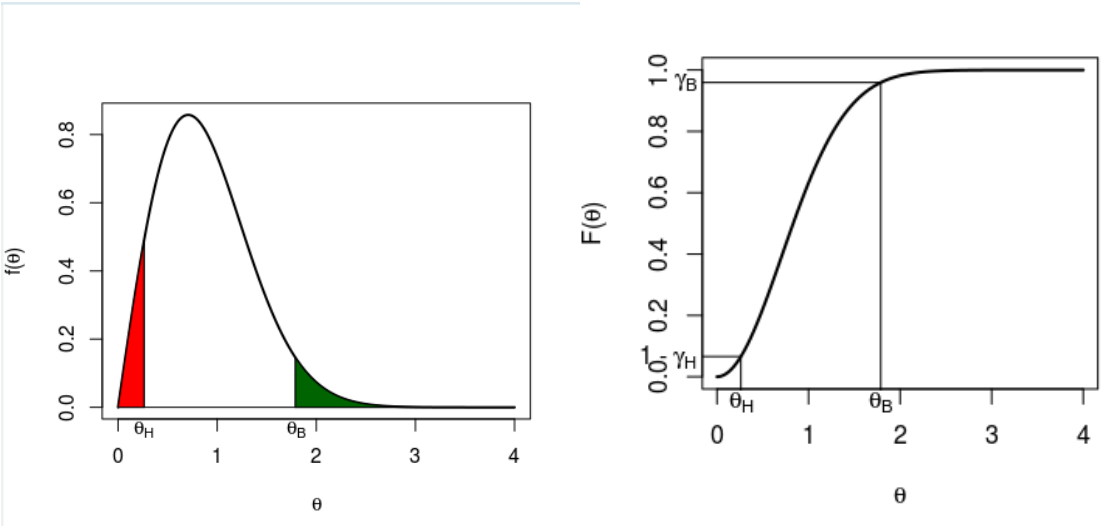


Рис. 6. Определение доверительных границ для параметра $\theta$ по плотности вероятности (слева) или функции распределения (справа). На графике слева площади закрашенные красным, белым и зелёным равны вероятностям $1-\gamma_H$ , $\gamma$ и $1-\gamma_B$ соответственно.

Оценки доверительных границ двухстороннего доверительного интервала определяются следующими зависимостями:

$$P\{\theta \ge \theta_H\} = \gamma_H , \qquad P\{\theta \le \theta_B\} = \gamma_B ,$$

$$\int_{-\infty}^{\theta_H} f(\theta)\, d\theta = 1-\gamma_H , \qquad \int_{\theta_B}^{\infty} f(\theta)\, d\theta = 1-\gamma_B ,$$

$$F(\theta_H) = 1-\gamma_H , \qquad F(\theta_B) = \gamma_B .$$

Из правого графика на Рис. 6 можно получить условие:

$$\gamma = \gamma_B - (1-\gamma_H) \quad \text{или} \quad \gamma_H + \gamma_B = 1 + \gamma .$$

При симметричной $f(\theta)$ для упрощения равенств выше величины $\gamma_H$ и $\gamma_B$ можно выбирать равными:

$$\gamma_H = \gamma_B = 0.5(1+\gamma) \ .$$

Не всегда удается построить плотность вероятности или функцию распределения для оценки параметра, поэтому для интервального оценивания некоторых параметров распределения СВ, например МО и дисперсии нормального распределения, удается подобрать специальные функции (*статистики*) $g$, которые зависят от результатов наблюдения $x_i, i=1,2,\ldots,n$ и от точечной оценки параметра $\theta$ : $g = \psi(\{x_i\},\theta)$ , для которых удается построить плотность вероятности $\varphi(g)$ . В этом случае для функции $g$ определяются доверительные границы $[G_H; G_B]$ , а оценки доверительных границ $[\theta_H; \theta_B]$ уже для параметра $\theta$ вычисляются как решения уравнений

$$\psi(\{x_i\},\theta_H) = G_H \ , \qquad \psi(\{x_i\},\theta_B) = G_B \ .$$

Для определения доверительных границ $\theta_H$ и $\theta_B$ некоторых параметров распределения СВ используются заранее подготовленные таблицы квантилей распределения функции $g$ в зависимости от значений $n$ и $\gamma$ . Для интервального оценивания математического ожидания и дисперсии нормальной СВ используются табличные значения квантилей распределений Стьюдента и Пирсона. Пользователям R нет нужды использовать таблицы, поскольку вычисления квантилей осуществляются функциями семейства **q*name*()**, где ***name*** — имя конкретного распределения в R, например для нормального распределения это функция ***qnorm()***.

Если плотность вероятности оценок параметра неизвестна, но можно полагать, что она симметрична и унимодальна, то доверительные границы можно определить непосредственно на основе несмещенной точечной оценки $\hat{\theta}$ и ее среднеквадратичного отклонения $\sigma_{\hat{\theta}}$ по зависимостям

$$\hat{\theta}_H = \hat{\theta} - u_{\gamma_H}\sigma_{\hat{\theta}} \ , \qquad \hat{\theta}_B = \hat{\theta} + u_{\gamma_B}\sigma_{\hat{\theta}} \ ,$$

где $u_{\gamma_H}$ и $u_{\gamma_B}$ это квантили нормированного центрированного распределения оценок, соответствующие вероятностям $\gamma_H$ и $\gamma_B$ . Зависимость обычно используется для интервального оценивания математического ожидания нормальной СВ с неизвестной дисперсией. При выборке ограниченного объема квантили определяются по распределению Стьюдента, а при большем объеме (как правило > 50-100) по нормальному распределению.

*Квантиль* — значение, которое СВ не превышает с заданной вероятностью, $\alpha$ -квантиль (т. е. квантиль порядка $\alpha \in (0;\ 1)$ ) это число $x_\alpha$ , такое что

$P[X \le x_\alpha] \le \alpha$ , в случае же непрерывного распределения $F(x_\alpha) = \alpha$ (см. Рис.7).

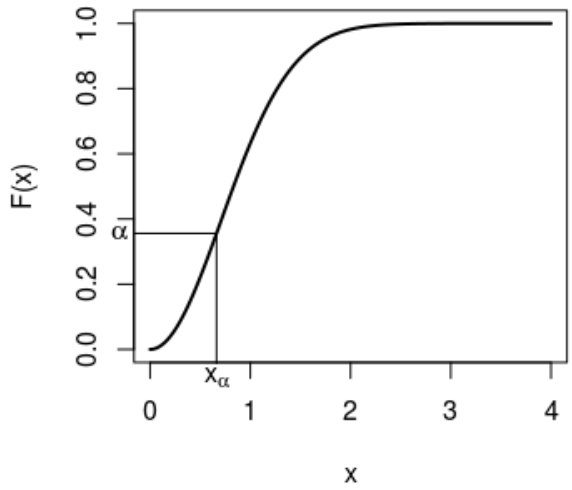


Рис. 7. Определение $\alpha$ -квантили.

## Оценивание первых моментов распределения

*Точечная оценка математического ожидания* для выборки $x_i, i = 1,2,..,n$ определяется как

$$\hat{M}_i = \frac{1}{n}\sum_{i=1}^{n} x_i \quad .$$

*Несмещенная оценка дисперсии*:

$$\hat{\sigma_X^2} = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \hat{M}_x\right)^2 \quad .$$

Центральная предельная теорема утверждает, что распределение оценки МО имеет асимптотически нормальное распределение при $x \to \infty$ с точечной оценкой МО в качестве МО и дисперсией

$$\sigma_{\hat{M}_x}^2 = \frac{\sigma_x^2}{n} \quad .$$

Интервальная оценка: $M \in [\hat{M}_X - u_p \sigma_{\hat{M}_X};\ \hat{M}_X + u_p \sigma_{\hat{M}_X}]$ , где $u_p$ - квантиль распределения Стьюдента.

## Проверка статистических гипотез

*Статистическая гипотеза* — предположение о типе неизвестного распределения и его параметрах. Проверяется на основе выборки реализаций СВ. Обычно используются следующие обозначения:

$H_0$ — нулевая гипотеза;

$H_1$ — конкурирующая (альтернативная) гипотеза, которая противоречит $H_0$.

Гипотеза может быть *простой* (одна точечная оценка параметра) или *сложной* (множество значений, интервал, бесконечно много). Выдвинутая гипотеза может быть верной или ложной, проверка производится статистическими методами, оперирующими случайными величинами и поэтому не обеспечивают полной достоверности принимаемых решений. Могут быть следующие варианты, при принятии решений:

1) выдвигается верная гипотеза, и она принимается;

2) выдвигается верная гипотеза, но она отклоняется;

3) выдвигается ложная гипотеза, и она отклоняется;

4) выдвигается ложная гипотеза, но она принимается.

*Ошибкой первого рода* называется принятие альтернативной гипотезы $H_1$, когда верна нулевая гипотеза $H_0$. *Ошибкой второго рода* называется принятие гипотезы $H_0$, когда верна альтернативная гипотеза $H_1$. Возможные ошибки характеризуются их вероятностями. Вероятность совершить ошибку первого рода называется *уровнем значимости* и обозначается как $\alpha$ . Вероятность совершить ошибку второго рода принято обозначать как $\beta$ . Вероятность отклонить нулевую гипотезу, когда верна альтернативная, называется *мощностью критерия*. Проверка гипотез осуществляется с помощью показателя $\Pi$ , зависящего от элементов выборки, а различение гипотез производиться благодаря тому, что априорные распределения показателя для $H_0$ и $H_1$ не совпадают. Правило принятия решения о верности той или иной гипотезы называют *статистическим критерием*, *статистическим тестом* или *критерием проверки статистической гипотезы*. На основе критерия определяется *критическая область*, попадание куда показателя $\Pi$ означает то, что нулевая гипотеза отвергается. Критическая область может быть *двухсторонней* или *односторонней*. Напротив, *область принятия гипотезы* это совокупность значений показателя $\Pi$ , при которых нулевая гипотеза принимается.

*Гипотеза о параметре распределения первого типа* — предположение, что неизвестный параметр $\theta$ равен $\theta_0$ (простая нулевая гипотеза) или $\theta \in [\theta_1; \theta_2]$ (сложная нулевая гипотеза). Если получена оценка $\hat{\theta}$ с плотностью вероятности $f(\theta)$ , то для проверки гипотезы задаются уровнем значимости $\alpha$ (вероятность ошибки первого рода), по выборке делается оценка доверительного интервала $[\hat{\theta}_H; \hat{\theta}_B]$ так что

$$P\left\{\theta \in [\hat{\theta}_H; \hat{\theta}_B]\right\} = 1-\alpha = 1-\alpha_1-\alpha_2 ,$$

где $\alpha_1 = 1 - \gamma_H$ , $\alpha_2 = 1 - \gamma_B$ . Для проверки простой гипотезы $H_0 : \theta = \theta_0$ используется *решающее правило*:

гипотеза принимается если $\theta_0 \in [\hat{\theta}_H ; \hat{\theta}_B]$ ;

гипотеза отклоняется если $\theta_0 \notin [\hat{\theta}_H ; \hat{\theta}_B]$ .

Для проверки сложной нулевой гипотезы $H_0 : \theta \in [\theta_1 ; \theta_2]$ применяется решающее правило:

гипотеза принимается если $[\hat{\theta}_H ; \hat{\theta}_B] \in [\theta_1 ; \theta_2]$ ;

гипотеза отклоняется если $\hat{\theta}_H > \theta_2$ или $\theta_B < \theta_1$ .

В случае же частичного перекрывания отрезков, принять решение нельзя и необходимо уточнить гипотезу или увеличить объем выборки. Один из пределов множества $[\theta_1 ; \theta_2]$ может быть бесконечным или иметь естественные пределы, в этом случае оценивается односторонний доверительный интервал. В таком случае проверка гипотезы осуществляется по правилу:

гипотеза принимается, если $\hat{\theta}_B \le \theta_2$ или $\hat{\theta}_H \ge \theta_1$ ;

гипотеза отклоняется в противном случае.

*Гипотеза о параметрах распределения второго типа* — предположение о том, что неизвестные параметры $\theta_1$ и $\theta_2$ в двух выборках совпадают, т. е. $H_0 : \theta_1 = \theta_2$ . Для проверки такой гипотезы на основании выборок делаются оценки $\hat{\theta}_1$ и $\hat{\theta}_2$ и выбирается показатель $\Pi = F(\hat{\theta}_1, \hat{\theta}_2)$ (часто это просто $\hat{\theta}_1 - \hat{\theta}_2$ или $\hat{\theta}_1 / \hat{\theta}_2$ ) и определяется его распределение как СВ. Для показателя вычисляется доверительный интервал при заданном уровне значимости. Решающее правило такое же как и для простой гипотезы первого типа, но для показателя $\Pi$ .

Если имеется возможность построить распределения оценок $\hat{\theta}_1$ и $\hat{\theta}_2$ , то можно непосредственно использовать вероятность неравенств $\theta_1 < \theta_2$ или $\theta_1 > \theta_2$ . Для примера рассмотрим вариант плотностей вероятностей оценок, построенных по выборкам и показанный на Рис. 8.

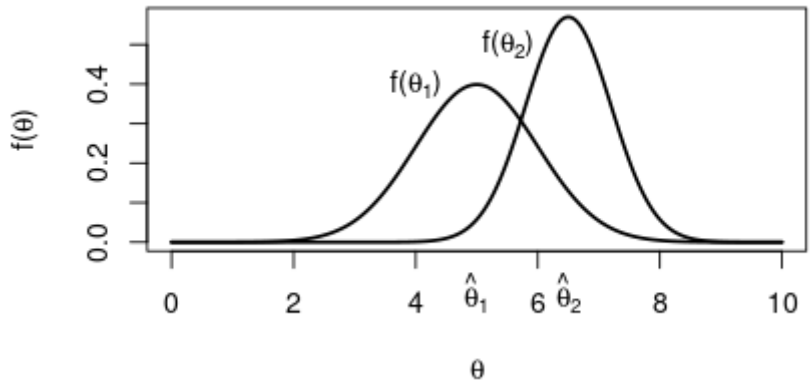

Рис.8. Пример плотности вероятности оценок

На основе плотностей вероятностей на Рис. 8 можно оценить вероятность того, что параметр $\theta_2$ во второй выборке будет больше или меньше параметра $\theta_1$ в первой выборке:

$$P\{\theta_2 > \theta_1\} = P_1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\hat{\theta}_2} f(\theta_1) f(\theta_2)\, d\theta_1\, d\theta_2 ,$$
$$P\{\theta_2 < \theta_1\} = P_2 = 1 - P_1 ,$$
$$P = max[P_1; P_2].$$

В данном случае при рассматриваемом уровне значимости $\alpha$ критическая область для показателя (вероятности) $P$ при проверке нулевой гипотезы $H_0: \theta_1 = \theta_2$ равна $P \in [1-\alpha;\ 1]$ . Поэтому решающее правило применяется следующее:

гипотеза принимается, если $P \leq' 1-\alpha$ ;

гипотеза отклоняется, если $P > 1-\alpha$ .

Гипотезы о распределении СВ, таким образом, делятся на два класса: гипотезы о совпадении распределений в двух выборках и гипотезы о типе распределения СВ. Для проверки гипотез используются специально подобранные показатели, отражающие степень близости распределений.

## Методы оценивания параметров распределения СВ и проверки статистических гипотез

Для точечного оценивания параметров распределения СВ на основе выборок используются такие классические методы математической статистики как метод максимального правдоподобия (ММП), метод моментов (ММ) и метод наименьших квадратов (МНК).

## Метод максимального правдоподобия

Для оценивания параметров распределения случайной величины по ММП необходимо знать тип распределения с точностью до неизвестного параметра $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$ . Если СВ дискретная, $x_q, q = 1,2,..,Q$ , то ее распределение описывается совокупностью вероятностей

$$P\{x_q \mid \boldsymbol{\theta}\} = p_q, \quad q = 1,2,\ldots,Q \ .$$

Если СВ непрерывная, то ее распределение можно представить в виде плотности вероятности или функции распределения:

$$f(x) = f(x, \boldsymbol{\theta}), \quad F(x) = F(x, \boldsymbol{\theta}) \ .$$

*Функция правдоподобия* это априорная вероятность получения совокупности возможных значений СВ в выборке $\{x_i\}$ , рассматриваемая как функция параметра $\boldsymbol{\theta}$ . Для дискретной величины функция правдоподобия это

$$L(\boldsymbol{\theta}) \ = \ P\left\{\{x_i\} \,\middle|\, \boldsymbol{\theta}\right\} \ = \ \prod_{i=1}^{n} P\{x_i \mid \boldsymbol{\theta}\} \ ,$$

для непрерывной —

$$L(\boldsymbol{\theta}) \ = \ f(\{x_i\}, \boldsymbol{\theta}) \ = \ \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta}) \ .$$

Сущность ММП в выборе параметра $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$ , доставляющего максимум функции правдоподобия

$$L(\boldsymbol{\theta})\big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \ = \ max$$

или

$$\ln L(\boldsymbol{\theta})\big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \ = \ max \ .$$

Если ФП имеет аналитический вид, то оценка параметров может быть сделана на основе решения системы уравнений (необходимых условий экстремума)

$$\left.\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j}\right|_{\hat{\theta}_j} \ = \ 0, \quad j = 1,2,\ldots,J$$

или

$$\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_j}\right|_{\hat{\theta}_j} = 0, \quad j=1,2,...,J \ .$$

Точность получаемых оценок характеризуется *ковариационной матрицей точечных оценок*

$$\boldsymbol{K}_{\hat{\boldsymbol{\theta}}} = \begin{pmatrix} \sigma^2_{\hat{\theta}_1} & K_{\hat{\theta}_1\hat{\theta}_2} & ... & K_{\hat{\theta}_1\hat{\theta}_J} \\ K_{\hat{\theta}_2\hat{\theta}_1} & \sigma^2_{\hat{\theta}_2} & ... & K_{\hat{\theta}_2\hat{\theta}_J} \\ ... & ... & ... & ... \\ K_{\hat{\theta}_J\hat{\theta}_1} & K_{\hat{\theta}_J\hat{\theta}_2} & ... & \sigma^2_{\hat{\theta}_J} \end{pmatrix},$$

где $\sigma^2_{\hat{\theta}_j}$ - *дисперсия j*-ой компоненты вектора оценки параметра $\hat{\theta}_j$ , а $K_{\hat{\theta}_j\hat{\theta}_k}$ - момент ковариации компонент вектора оценки $\hat{\theta}_j$ и $\hat{\theta}_k$ . Аналитический вид матрицы $\boldsymbol{K}_{\hat{\boldsymbol{\theta}}}$ можно представить как обращение *информационной матрицы Фишера* при полученных точечных оценках

$$\boldsymbol{K}_{\hat{\boldsymbol{\theta}}} = -\left\{\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}\right\}^{-1}_{\hat{\boldsymbol{\theta}}} \ .$$

Если $\theta$ скаляр, а не вектор, то дисперсия оценки определяется по более простой формуле

$$\sigma^2_{\hat{\theta}} = -\left(\frac{d^2 \ln L(\theta)}{d\,\theta^2}\right)^{-1}_{\hat{\theta}} \ .$$

Доказано, что оценки ММП в общем случае являются асимптотически несмещенными и асимптотически эффективными (при объеме выборки $n \to \infty$ ).

## Метод моментов

Рассмотрим СВ *X* с плотностью вероятности

$$f(x) = f(x, \boldsymbol{\theta}), \qquad \boldsymbol{\theta} = (\theta_{1,}\theta_{2,...}, \theta_J) \ .$$

Для оценивания $\boldsymbol{\theta}$ надо получить *J* первых моментов. Для примера без потери общности положим $J = 2$ , т. е.

$$f(x) = f(x, \theta_1, \theta_2) \ .$$

Два первых момента это математическое ожидание и дисперсия,

$$M_X = \int_{\Omega_x} x f(x, \theta_1, \theta_2) dx = F_1(\theta_1, \theta_2) ,$$

$$D_X = \int_{\Omega_x} (x - M_X)^2 f(x, \theta_1, \theta_2) dx = F_2(\theta_1, \theta_2) .$$

Если $\theta_1$ и $\theta_2$ неизвестны, но получена выборка $x_i, i = 1,2,\ldots,n$ , то по ней можно сделать оценки $\hat{M}_X$ и $\hat{D}_X$ , дисперсии которых

$$\sigma^2_{\hat{M}_x} = \frac{\hat{D}_X}{n}, \qquad \sigma^2_{\hat{D}_x} = \frac{2\hat{D}^2_X}{n} .$$

Точечные оценки параметров распределения $\theta_1$ и $\theta_2$ вычисляются путем *приравнивания выражений для моментов распределения их точечным оценкам*. Т.е. для случая $J = 2$ нужно решить следующую систему уравнений:

$$\begin{cases} F_1(\hat{\theta}_1, \hat{\theta}_2) = \hat{M}_X \\ F_2(\hat{\theta}_1, \hat{\theta}_2) = \hat{D}_X \end{cases} .$$

Решение этой системы выполняется различными методами в зависимости от сложности функций $F_1$ и $F_2$. Иногда в простых случаях можно получить явные выражения вида

$$\theta_j = \varphi_j(\hat{M}_X, \hat{D}_X), \quad j = 1,2 .$$

Дисперсии оценок с использованием линеаризации определяются как

$$\sigma^2_{\hat{\theta}_j} = \left[\frac{\partial \varphi_j(\hat{M}_X, \hat{D}_X)}{\partial \hat{M}_X}\right]^2 \sigma^2_{\hat{M}_x} + \left[\frac{\partial \varphi_j(\hat{M}_X, \hat{D}_X)}{\partial \hat{D}_X}\right]^2 \sigma^2_{\hat{D}_x}, \quad j = 1,2 .$$

Пример: Оценивание параметров по ММ равномерного распределения

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

Для получения оценок параметров *A* и *B* нужно решить систему

$$\begin{cases} \frac{1}{2}(\hat{A}+\hat{B}) = \hat{M}_X \\ \frac{(\hat{B}-\hat{A})^2}{12} = \hat{D}_X \end{cases},$$

решение которой

$$\begin{aligned} \hat{A} &= \hat{M}_X - \sqrt{3\hat{D}_X}, \\ \hat{B} &= \hat{M}_X + \sqrt{3\hat{D}_X} \end{aligned} .$$

Выражения для дисперсий оценок параметров *A* и *B* имеют вид

$$\sigma^2_{\hat{A}} = \sigma^2_{\hat{B}} = \sigma^2_{\hat{M}_x} + \frac{3\sigma^2_{\hat{D}_x}}{4\hat{D}_X} .$$

Нормальное распределение имеет два параметра: математическое ожидание и дисперсию. Поэтому нет нужды решать систему уравнение по ММ, поскольку оценки ММ совпадают с точечными оценками по выборке,

$$\hat{M}_X = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma^2_{\hat{M}_x} = \frac{\hat{D}_X}{n};$$

$$\hat{D}_X = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{M}_X)^2, \qquad \sigma^2_{\hat{D}_x} = \frac{2\hat{D}_X^2}{n} .$$

## Метод наименьших квадратов

Метод наименьших квадратов является распространенным методом оценивания параметров любых параметрических моделей, не обязательно вероятностных распределений. Его можно использовать для оценки первых двух моментов по стабильной или динамичной выборке. МНК можно использовать для решения уравнений, связанных с ММП и ММ. МНК так же используется для оценивания динамики неслучайных параметров и функций по результатам измерений, искаженных случайными погрешностями, а также для поиска зависимостей исследуемых показателей от различных измеряемых параметров и факторов.

В качестве примера рассмотрим применение МНК для подбора параметров модели описывающей динамику математического ожидания:

$$M(t) = F(t,\boldsymbol{\theta}) ,$$

где $F()$ это известная функция или алгоритм, а $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_J)$ - вектор неизвестных параметров. Наблюдения $x_i, i = 1,2,..,n$ в моменты времени $t_1, t_2, ..., t_n$ интерпретируются как

$$x_i \;=\; F(t_i, \boldsymbol{\theta}) + \delta_i \;\;,$$

где $\delta_i$ - случайные флуктуации, как правило с нулевым средним нормальным распределением. Собственно МНК в общем виде это решение оптимизационной задачи вида

$$\min_{\boldsymbol{\theta}} \;\sum_{i=1}^{n} \left(x_i - F(t_i, \boldsymbol{\theta})\right)^2 \;\;,$$

заключающейся в поиске параметров $\boldsymbol{\theta}$ , которые доставляют минимум сумме квадратов разностей между измеряемой величиной и функцией ее моделирующей. В общем случае задача может оказаться сложной и не иметь аналитического решения. Задача может оказаться плохо обусловленной и требовать хорошего априорного приближения вектора параметров $\boldsymbol{\theta_0}$ .

Для упрощения решения задачи, модель может представляться в *линеаризованном* виде:

$$M(t) \;\approx\; F(t, \boldsymbol{\theta_0}) + \sum_{j=1}^{J} \left.\frac{\partial F(t, \boldsymbol{\theta})}{\partial \theta_j}\right|_{\boldsymbol{\theta_0}} (\theta_j - \theta_{j0}) \;=\; F(t, \boldsymbol{\theta_0}) + \sum_{j=1}^{J} \varphi_j(t) \Delta\theta_j \;\;.$$

Модель наблюдений уже записывается в линеаризованном виде следующим образом:

$$\sum_{j=1}^{J} \varphi_j(t_i) \Delta\theta_j + \delta_i \;=\; \Delta x_i \;, \qquad i = 1,2,..,n \;\;,$$

где $\Delta x_i = x_i - F(t_i, \boldsymbol{\theta_0})$ . Собственно в данной записи линеаризованная модель это система линейных уравнений.

Линеаризованную модель можно записать в вектор-матричной форме как

$$\boldsymbol{\Phi} \cdot \boldsymbol{\Delta\Theta} + \boldsymbol{\Delta} \;=\; \boldsymbol{\Delta X} \;\;,$$

где $\boldsymbol{\Delta\theta} = \{\theta_j\}$ - вектор отклонений неизвестных параметров от начального приближения, $\boldsymbol{\Phi} = \{\varphi_j(t_i)\}$ - матрица преобразования, $\boldsymbol{\Delta}$ - вектор случайных флуктуаций, $\boldsymbol{\Delta X}$ - вектор отклонений данных наблюдений от начальных значений модели.

Если известна ковариационная матрица $\boldsymbol{K}_{\Delta}$ случайного вектора $\boldsymbol{\Delta}$, то *обобщенный метод наименьших квадратов* предполагает минимизацию следующей целевой функции:

$$S_{об} = \boldsymbol{V}^{\boldsymbol{T}} \boldsymbol{K}_{\Delta}^{-1} \boldsymbol{V} = \min_{\boldsymbol{\theta}},$$

где $\boldsymbol{V} = \left(x_1 - F(\boldsymbol{\theta}, t_1),\ x_2 - F(\boldsymbol{\theta}, t_2),\ \ldots,\ x_n - F(\boldsymbol{\theta}, t_n)\right)^T$ - вектор отклонений.

Если флуктуации не коррелированы и имеют разные дисперсии, то используется взвешенный метод наименьших квадратов, предполагающий минимизацию следующей целевой функции:

$$S_{взв} = \sum_{i=1}^{n} \sigma_i^2 \left[x_i - F(\boldsymbol{\theta}, t_i)\right]^2 = min \ .$$

И наконец, если флуктуации не коррелированы и имеют постоянную дисперсию, т. е. $\sigma_i = const = \sigma,\ \forall i$, то применяется *классический* МНК с целевой функцией вида

$$S_{кл} = \sum_{i=1}^{n} \left[x_i - F(\boldsymbol{\theta}, t_i)\right]^2 = min \ .$$

Для линейной или линеаризованной модели наблюдений аналогичные целевые функции легко получить подстановкой вместо $F(\boldsymbol{\theta}, t)$ её линеаризованной версии. Преимущество линейного варианта заключается в том, что оценки отклонений параметра $\boldsymbol{\Delta}\hat{\boldsymbol{\theta}}$ линейно зависят от $\boldsymbol{\Delta X}$, т. е. $\boldsymbol{\Delta}\hat{\boldsymbol{\theta}} = \boldsymbol{B}\boldsymbol{\Delta X}$, где $\boldsymbol{B}$ это некоторая матрица преобразования и в этом случае легко получить ковариационную матрицу оценки векторного параметра:

$$\boldsymbol{K}_{\boldsymbol{\Delta}\hat{\boldsymbol{\theta}}} = \boldsymbol{B}\boldsymbol{K}_{\Delta}\boldsymbol{B}^{\boldsymbol{T}} \ .$$

Для линеаризованной модели наблюдений для вычисления точечных оценок и ковариационной матрицы оценок вектора параметров готовые выражения имеют следующий вид:

для обобщенного МНК

$$\boldsymbol{\Delta}\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Phi}^{\boldsymbol{T}} \boldsymbol{K}_{\Delta}^{-1} \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\boldsymbol{T}} \boldsymbol{K}_{\Delta}^{-1} \boldsymbol{\Delta X}, \qquad \boldsymbol{K}_{\boldsymbol{\Delta}\hat{\boldsymbol{\theta}}} = \left(\boldsymbol{\Phi}^{\boldsymbol{T}} \boldsymbol{K}_{\Delta}^{-1} \boldsymbol{\Phi}\right)^{-1} \ ;$$

для взвешенного МНК

$$\Delta\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Phi}_{\boldsymbol{н}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{н}}\right)^{-1}\boldsymbol{\Phi}_{\boldsymbol{н}}^{\boldsymbol{T}}\Delta\boldsymbol{X}_{\boldsymbol{н}}, \qquad \boldsymbol{K}_{\Delta\hat{\boldsymbol{\theta}}} = \left(\boldsymbol{\Phi}_{\boldsymbol{н}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{н}}\right)^{-1} \ ;$$

для классического МНК

$$\Delta\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\boldsymbol{T}}\Delta\boldsymbol{X}, \qquad \boldsymbol{K}_{\Delta\hat{\boldsymbol{\theta}}} = \sigma^2\left(\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi}\right)^{-1} \ .$$

Здесь в формулах для взвешенного МНК использованы взвешенные значения вектора и матрицы, т. е. $\Delta\boldsymbol{X} = \left(\Delta x_i/\sigma_i\right)$, $\boldsymbol{\Phi}_{\boldsymbol{н}} = \left(\varphi_j(t_i)/\sigma_i\right)$ . Если в случае классического МНК дисперсия флуктуаций $\sigma^2$ неизвестна, то *несмещенную точечную оценку дисперсии флуктуаций* можно определить по остаточным невязкам как

$$\hat{\sigma^2} \ = \ \frac{1}{n-J}\left(\Delta\boldsymbol{X}-\boldsymbol{\Phi}\,\Delta\hat{\boldsymbol{\theta}}\right)^T\left(\Delta\boldsymbol{X}-\boldsymbol{\Phi}\,\Delta\hat{\boldsymbol{\theta}}\right) \ .$$

На основе оценок вектора отклонений вектора параметров от начального приближения определяется оценка самого вектора и его ковариационной матрицы:

$$\hat{\boldsymbol{\theta}} \ = \ \boldsymbol{\theta_0} + \Delta\hat{\boldsymbol{\theta}} \ ; \qquad \boldsymbol{K}_{\hat{\boldsymbol{\theta}}} \ = \ \boldsymbol{K}_{\Delta\hat{\boldsymbol{\theta}}} \ .$$

## Регрессионный анализ

При анализе данных наблюдений могут решаться задачи исследования зависимости параметров распределения СВ от различных факторов. Для таких задач используются методы *регрессионного анализа*.

Рассмотрим зависимость вида

$$M_X(\boldsymbol{g}) \ = \ a_0+a_1g_1+a_2g_2+\ldots+a_Jg_J = a_0+\sum_{j=1}^{J}a_jg_j \ = \ a_0+\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{g} \ .$$

Для простоты записи можно полагать, что всегда $g_0 = 1$ и тогда запись упростится, т. е.

$$M_X(\boldsymbol{g}) \ = \ \boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{g} \ ,$$

где векторы $\boldsymbol{a} = (a_0, a_1, a_2, \ldots, a_J)^T$, $\boldsymbol{g} = (1, g_1, g_2, \ldots, g_J)^T$ , размерность которых *J+1*.

Если от какого-то фактора имеется нелинейная зависимость, например

$$M_X(z) = b_1 z + b_2 z^2 \ ,$$

то $z$ и $z_2$ представляются как отдельные факторы, т. е. $z=g_k$, $z^2=g_{k+1}$ и можно снова пользоваться аппаратом линейной регрессии.

Для оценивания коэффициентов регрессии $a_0, a_1, \ldots, a_J$ проводятся наблюдения с выборкой $x_i, i=1,2,\ldots,n$ при различных значениях факторов $g_1, g_2, \ldots, g_J$ (т. е. вместе с выборкой нужно получить набор $\boldsymbol{g_i}$, $i=1,2,\ldots,n$ ) и используется следующая модель наблюдений:

$$M_X(\boldsymbol{g}) + \delta_i = x_i, \quad i=1,\ldots,n$$

Задача регрессионного анализа сводится к МНК для решения задачи нахождения минимума (в пространстве коэффициентов регрессии)

$$\min_{\boldsymbol{a}} \sum_{i=1}^{n} \left( x_i - \sum_{j=0}^{J} a_j g_{ji} \right)^2 \ ,$$

здесь надо помнить о соглашении $g_{0i} = 1, \ \forall i$ . Задача может быть переформулирована в векторно-матричном виде как

$$\min_{\boldsymbol{a}} (\boldsymbol{x} - \boldsymbol{G}\,\boldsymbol{a})^T (\boldsymbol{x} - \boldsymbol{G}\,\boldsymbol{a}) \ .$$

В соответствии с классическим МНК можно получить зависимости для *точечных оценок вектора регрессионных коэффициентов и их ковариационной матрицы*:

$$\hat{\boldsymbol{a}} = (\boldsymbol{G}^{\boldsymbol{T}} \boldsymbol{G})^{-1} \boldsymbol{G}^{\boldsymbol{T}} \boldsymbol{x} \ ;$$

$$\hat{\sigma^2} = \frac{1}{n-J} (\boldsymbol{x} - \boldsymbol{G}\,\hat{\boldsymbol{a}})^T (\boldsymbol{x} - \boldsymbol{G}\,\hat{\boldsymbol{a}}) \ ;$$

$$\boldsymbol{K}_{\hat{\boldsymbol{a}}} = \hat{\sigma^2} (\boldsymbol{G}^{\boldsymbol{T}} \boldsymbol{G})^{-1} \ .$$

Все эти вычисления возможны только если все факторы линейно независимы, если какие-то факторы линейно зависимы, то необходимо уменьшить число этих факторов, оставив только линейно независимые или функции от них.

*Упражнение. Выполнить работу предложенную по ссылке*
http://environmentalcomputing.net/linear-regression/

## Методы интервального оценивания параметров распределения СВ

### Оперативный метод

Если для неизвестного параметра $\theta$ получена точечная оценка $\hat{\theta}$ и ее среднеквадратичное отклонение $\sigma_{\hat{\theta}}$, то оценка границ доверительного интервала $[\theta_H; \theta_B]$ при заданной доверительной вероятности $\gamma$

$$\hat{\theta}_H = \hat{\theta} - u_{\gamma_н}\sigma_{\hat{\theta}} ,$$

$$\hat{\theta}_B = \hat{\theta} + u_{\gamma_в}\sigma_{\hat{\theta}} ,$$

где $\gamma_H = P\{\theta \geq \theta_H\}$, $\gamma_B = P\{\theta \leq \theta_B\}$, $\gamma_H + \gamma_B = 1 + \gamma$, $u_{\gamma_н}, u_{\gamma_в}$ - квантили распределения точечной оценки, соответствующие вероятностям $\gamma_H$ и $\gamma_B$ соответственно. Если при решении практических задач предполагают нормальное распределение оценки параметра $\hat{\theta}$, то квантили распределения определяются по формуле Лапласа. В R это qnorm( $\gamma$ ,0,1), т. е.

$$u_\gamma = qnorm(\gamma, 0, 1) .$$

Если распределение $\hat{\theta}$ отличается от нормального, но известно, что оно симметрично и унимодально, то квантили больше

$$u_\gamma = \frac{\sqrt{2}}{3\sqrt{1-\gamma}} .$$

Если о распределении вообще ничего не известно, то квантили еще больше

$$u_\gamma = \frac{1}{\sqrt{2(1-\gamma)}} .$$

***Упражнение в R:*** *Пусть* $\hat{M}_X = 3.0,\ \hat{\sigma}_{\hat{M}_X} = 0.45,\ \gamma = 0.9$ . *Определить* $[\theta_H; \theta_B]$ *для всех вышеописанных случаев.*

### Интервальное оценивание МО СВ с нормальным распределением и неизвестной дисперсией

В этом случае используется показатель (статистика)

$$t = \frac{\hat{M}_X - M_X}{\sigma_{\hat{M}_x}}$$

имеющий распределение Стьюдента с числом степеней свободы $k = n - 1$ , где $n$ это объем выборки. (см. презентацию «*Распределения дискретных и непрерывных случайных величин и знакомство с ними в R*»). Оценка границ доверительного интервала выполняется по формулам

$$\hat{M}_{XH} = \hat{M} - t(\gamma, k)\sigma_{\hat{M}_x} ,$$

$$\hat{M}_{XB} = \hat{M} + t(\gamma, k)\sigma_{\hat{M}_x} ,$$

где $t(\gamma, k)$ - квантиль распределения Стьюдента, которая в R вычисляется с помощью функции $qt(\gamma, k, ncp = 0)$ .

## Связь интервальных и точечных оценок

Распределение Стьюдента получено для случайной величины $t$ при фиксированном значении $M_X$, для случайной величины $X$, имеющей нормальное распределение. Если распределение Стьюдента верно отражает распределение $t$, то можно построить распределение $\hat{M}_X$ . Обозначим неизвестную величину $M_X$ как $m$, тогда

$$t = \frac{\hat{M}_X - m}{\sigma_{\hat{M}_x}} .$$

Тогда плотность распределения $m$ может быть выражена через распределение Стьюдента как

$$g(m) = f\left(\frac{\hat{M}_x - m}{\sigma_{\hat{M}_x}}\right) \cdot \left|\frac{dt}{dm}\right| = \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\,\Gamma(k/2)} \cdot \left[1 + k^{-1}\left(\frac{\hat{M}_X - m}{\sigma_{\hat{M}_x}}\right)^2\right]^{-\frac{k+1}{2}} \cdot \left(\frac{1}{\sigma_{\hat{M}_x}}\right)$$

Интегрирование этой плотности приводит к тем же оценкам границ доверительного интервала $[\hat{M}_{XH}; \hat{M}_{XB}]$ , а также можно воспользоваться равенствами

$$\hat{M}_x = \int_{-\infty}^{\infty} m\, g(m)\, dm ,$$

$$\sigma^2_{\hat{M}_x} = \int_{-\infty}^{\infty} (m - \hat{M}_X)^2\, g(m)\, dm .$$

Оценка математического ожидания совпадает с точечной оценкой по выборке, т. е. с $\frac{1}{n}\sum_{i=1}^{n} x_i$ , однако дисперсия отличается от дисперсии определяемой по ММП.

## Определение доверительных границ для дисперсии

Рассмотрим случайную величину *X* с неизвестными $M_X$ и $D_X$ . Пусть имеется выборка $\{x_i\}$ объема *n*. Используется в данном случае показатель (статистика) следующего вида:

$$\chi^2 = \frac{n\hat{D}_X}{D_x} .$$

При нормальном распределении случайной величины *X* показатель $\chi^2$ имеет $\chi^2$ -распределение с числом степеней свободы $k = n-1$ . Для удобства записи введем обозначение $y = \chi^2 \in [0; +\infty]$ , тогда функция плотности вероятности имеет вид

$$f(y) = \left[\Gamma(k/2)\cdot 2^{k/2}\right]^{-1}\cdot y^{(k/2-1)}\cdot e^{-y/2} ,$$

в R это функция `dchisq()`.

Для оценки границ доверительного интервала задаются вероятностями $\gamma_H$ и $\gamma_B$ так чтобы

$$P\{D \le D_B\} = 1 - \gamma_H ,$$

$$P\{D \ge D_H\} = \gamma_B .$$

Затем определяются квантили $\chi_1^2(\gamma_H, k)$ и $\chi_2^2(\gamma_B, k)$ и определяются доверительные границы по формулам

$$\hat{D_{XH}} = \frac{n\hat{D}_X}{\chi_2^2(\gamma_B, k)} ,$$

$$\hat{D_{XB}} = \frac{n\hat{D}_X}{\chi_1^2(\gamma_H, k)} .$$

Пример: $n = 10$ , $\hat{D}_X = 2.03$ , $\gamma_H = 0.05$ , $\gamma_B = 0.95$ . Можно получить для этого случая

$$\hat{D_{XH}} \approx 1.20, \qquad \hat{D_{XB}} \approx 6.11 \ .$$

*Упражнение в R: получить ответы примера самостоятельно используя функцию* `qchisq()`.

### Связь интервальных и точечных оценок

Построим на основе $\chi^2$-распределения для статистики $\chi^2$ распределение для $D_X$ .

$$y = \frac{n\hat{D}_X}{d} \ ,$$

$$g(d) = f\left(\frac{n\hat{D}_X}{d}\right)\left|\frac{\partial y}{\partial d}\right| = \left[\Gamma(k/2)\,2^{k/2}\right]^{-1}\left(\frac{n\hat{D}_X}{d}\right)^{k/2-1} e^{-\frac{n\hat{D}_X}{2d}}\cdot\left(\frac{n\hat{D}_X}{d^2}\right) \ .$$

В правой части все выражение, за исключением последней скобки вычисляется в R функцией `dchisq(y,k,…)`. На основе этой плотности можно вычислить несмещенную точечную оценку дисперсии и ее дисперсию,

$$\hat{D} = \int_0^\infty d\cdot g(d)\cdot dd; \qquad \sigma^2_{\hat{D}} = \int_0^\infty (d-\hat{D})^2 g(d)\,dd \ .$$

Эти оценки не совпадают с оценками ММП.

## Методы проверки статистических гипотез

### Проверка гипотезы о математическом ожидании случайной величины

Пусть наша нулевая гипотеза заключается в том, что математическое ожидание случайной величины $\underline{X}$ совпадает с некоторым значением $M_0$ , т. е.

$$H_0: M_X = M_0 \ ;$$
$$H_1: M_X \neq M_0 \ .$$

Проверка гипотезы осуществляется следующим образом. Задается уровень значимости $\alpha$ , вычисляются выборочные оценки

$$\hat{M}_X = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad \sigma^2_{\hat{M}_x} = \frac{\sigma^2_X}{n} \ .$$

Если дисперсия $\sigma^2_X$ неизвестна, то делается и ее выборочная оценками

$$\hat{\sigma}^2_X = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \hat{M}_X\right)^2 \ .$$

Для проверки гипотезы используется следующий показатель (статистика):

$$z = \frac{\hat{M}_X - M_0}{\sigma_{\hat{M}_x}} \ .$$

В зависимости от того известна ли $\sigma^2_X$ , используются два способа.

*1-ый способ, $\sigma^2_X$ известна.* В этом случае показатель описывается стандартным нормальным распределением с нулевым средним и единичной дисперсией, т. е. $z \in N[0,1]$ . Определяется критическое значение показателя $z_{кр}$ на основе функции распределения стандартного нормального распределения (с нулевым средним и единичной дисперсией) с использованием функции Лапласа из соотношения

$$\Phi(z_{кр}) = 1 - \alpha \ .$$

В учебной литературе по математической статистике имеется разночтение, что именно считать функцией Лапласа, это либо $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}} dz$ , либо $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x e^{-\frac{z^2}{2}} dz$ (выражения отличаются нижним пределом интегрирования), поэтому уравнение для определения $z_{кр}$ может быть записано как

$$0.5 + \Phi(z_{кр}) = 1 - \alpha \ ,$$

но в данном курсе мы будем считать, что функция $\Phi()$ это функция распределения стандартного нормального распределения. Таким образом $z_{кр}$ можно определить функцию вычисления квантилей в R

$$z_{кр} = qnorm(1-\alpha, 0, 1) \ .$$

По выборке вычисляется показатель $\hat{z}$ и сравнивается с $z_{кр}$ . Для проверки гипотезы о математическом ожидании случайной величины используется следующее решающее правило:

гипотеза $H_0: M_X = M_0$ принимается, если $|\hat{z}| \le z_{кр}$ ,

гипотеза $H_0$ отклоняется, если $|\hat{z}| > z_{кр}$ .

*2-ой способ,* $\sigma_X^2$ *неизвестна.* Вместо дисперсии используется ее выборочная оценка $\hat{\sigma}_X^2$ и вычисляются уже квантили распределения Стьюдента с $k = n-1$ числом степеней свободы, т. е. в R это

$$z_{кр} = qt(1-\alpha, n-1) .$$

Правило проверки гипотезы то же, что и для 1-го способа.

Следует отметить, что назначение слишком низкого уровня значимости может привести к отбрасыванию альтернативной гипотезы, когда она все таки верна. Поэтому назначение низкого уровня значимости следует делать тогда, когда есть высокая уверенность в верности нулевой гипотезы. В противном случае при малом уровне $\alpha$ с большой вероятностью $1-\alpha$ может быть отклонена верная альтернативная гипотеза.

## Поверка гипотез о дисперсии случайной величины

Пусть имеется выборка $x_i, i = 1,2,..,n$ из генеральной совокупности с неизвестными математическим ожиданием $M_X$ и дисперсией $D_X$ . Проверяется гипотеза о том, что дисперсия данной СВ не превышает некоторого заданного значения $D_0$ , т. е. проверяются гипотезы

$$H_0: D_X \le D_0 ,$$

$$H_1: D_X > D_0 .$$

Если СВ имеет нормальное распределение, то проверка $H_0$ делается по критерию $\chi^2$ К.Пирсона

$$\chi^2 = (n-1)\frac{\hat{\sigma}_X^2}{D_0} ,$$

где $\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{M}_X)^2$ , $\hat{M}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$ . Задается $\alpha$ - уровень значимости, при нормальном распределении СВ, показатель имеет $\chi^2$ -

распределение с $k=n-1$ числом степеней свободы. Затем находится квантиль $\chi^2_{кр}$ такой что

$$P\{\chi^2 > \chi^2_{кр}\} = \alpha ,$$

в R это делается легко $\chi^2_{кр} = qchisq(1-\alpha, k)$ . Применяется решающее правило:

гипотеза $H_0 : D_X \le D_0$ принимается, если $\chi^2 \le \chi^2_{кр}$ ;

гипотеза $H_0$ отклоняется если $\chi^2 > \chi^2_{кр}$ .

Принятие нулевой гипотезы лишь означает, что данные наблюдений не противоречат ей. При назначении $\alpha$ следует учитывать, что вероятность отклонить альтернативную гипотезу, если она верна, равна $1-\alpha$ .

При проверке гипотез полезно определить вероятность того, что при справедливой нулевой гипотезе значение показателя $\chi^2$ может превышать полученную реализацию $\hat{\chi}^2$ . Такая вероятность определяется на основе выражения

$$P\{\chi^2 > \hat{\chi}^2\} = 1 - \gamma(\hat{\chi}^2) ,$$

где $\gamma(\hat{\chi}^2)$ - вероятность, определяемая по функции $\chi^2$ -распределения, в R это делается с помощью функции $pchisq(\hat{\chi}^2, k)$ . Эта величина показывает вероятность того, что при справедливости нулевой гипотезы оценка дисперсии по выборке могла быть еще выше.

## Проверка гипотезы о математических ожиданиях случайных величин по двум выборкам

Пусть имеется две выборки $\{x_i\}$ и $\{y_i\}$ объемами $n_1$ и $n_2$ соответственно. Рассмотрим гипотезу о равенстве $M_X$ и $M_Y$ для случая нормального распределения случайных величин в обоих выборках. Для проверки гипотезы используются точечные оценки

;

$$\hat{M}_Y = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i \; ; \qquad \sigma^2_{\hat{M}_Y} = \frac{\sigma^2_Y}{n_2} .$$

Используется показатель $z = \frac{\hat{M}_X - \hat{M}_Y}{\sigma_{\Delta M}}$ , где $\sigma_{\Delta M}$ - среднеквадратическое отклонение разности разности оценок математических ожиданий:

$$\sigma_{\Delta M} = \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}} \ .$$

Способ проверки зависит от того, известны ли дисперсии СВ или нет.

*1-ый способ, дисперсии* $\sigma_X^2$ *и* $\sigma_Y^2$ *известны.* В этом случае при нормально распределенных случайных величинах, показатель *z* имеет стандартное нормальное распределение (с нулевым средним и единичной дисперсией), поэтому для проверки гипотезы задается уровень значимости $\alpha$, определяется $z_{кр}$ (с использованием R)

$$z_{кр} = qnorm(1-\alpha, 0, 1)$$

и используется решающее правило

гипотеза $H_0 : M_X = M_Y$ принимается, если $|\hat{z}| \le z_{кр}$ ;
гипотеза $H_0$ отклоняется, если $|\hat{z}| > z_{кр}$ .

*2-ой способ, дисперсии* $\sigma_X^2$ *и* $\sigma_Y^2$ *известны.* Вместо известных дисперсий в этом случае используются их оценки

$$\hat{\sigma_X^2} = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(x_i - \hat{M}_X\right)^2 \ ; \qquad \hat{\sigma_Y^2} = \frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(y_i - \hat{M}_Y\right)^2 \ ,$$

а проверка гипотезы проводиться по *t*-критерию Стьюдента

$$t = \frac{|\hat{M}_X - \hat{M}_Y|}{\hat{\sigma_{\Delta M}}} \ , \text{где} \quad \hat{\sigma_{\Delta M}} = \sqrt{\frac{\hat{\sigma_X^2}}{n_1} + \frac{\hat{\sigma_Y^2}}{n_2}} \ .$$

Определяется критическое значение критерия, в R используется функция qt():

$$t_{кр} = qt(1-\alpha, n_1+n_2-2) \ .$$

Решающее правило в этом случае:

гипотеза $H_0 : M_X = M_Y$ принимается, если $|\hat{t}| \le t_{кр}$ ;
гипотеза $H_0$ отклоняется, если $|\hat{t}| > t_{кр}$ .

## Проверка гипотез о дисперсиях СВ по двум выборкам

Итак, имеется две выборки $x_i, i=1,2,..,n_1$ и $y_i, i=1,2,...,n_2$ и нужно проверить гипотезу о равенстве дисперсий, т. е. $H_0: \sigma_X^2 = \sigma_Y^2$ . Используется *F*-критерий Фишера-Снедекора

$$F = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} ,$$

который при нормальных СВ и справедливости нулевой гипотезы имеет *F*-распределение Фишера-Снедекора, которое зависит от двух чисел степеней свободы $k_1=n_1-1$ и $k_2=n_2-1$ . Выбирается критическое значение *F* из условия

$$P\{ F > F_{кр} \} = \alpha ,$$

что в среде R сделать просто с помощью функции `qf()`

$$F_{кр} = qf(1-\alpha, k_1, k_2, ncp=0) .$$

И наконец, применяется решающее правило:

гипотеза $H_0: \sigma_X^2 = \sigma_Y^2$ принимается, если $\hat{F} \le F_{кр}$ ;
гипотеза отклоняется, если $\hat{F} > F_{кр}$ .

## Проверка гипотезы об однородности распределений СВ в двух выборках

Гипотезу можно сформулировать как $H_0: f(x) = f(y)$ . Наиболее мощным для этого случая считается критерий Вилкоксона (Wilcoxon), который применяется к упорядоченным в виде вариационного ряда выборкам

$$\{x_i\} = x_1, x_2, ..., x_{n_1}, \quad x_1 \le x_2 \le x_3 \le ... \le x_{n_1} ,$$

$$\{y_i\} = y_1, y_2, ..., y_{n_2}, \quad y_1 \le y_2 \le y_3 \le ... \le y_{n_2} .$$

Выборки объединяются в общий вариационный ряд, например такой

$$\{x, y\} = x_1, x_2, y_1, x_3, y_2, ..., y_{n_2}, x_{n_1} ,$$

элементы которого нумеруются сквозным образом, а номер (или ранг) перечисляет элементы общего вариационного ряда

$$r_q = 1,2,3,4,5,\ldots,n-1,n \ ,$$

где $n = n_1 + n_2$ . Вводятся показатели следующего вида

$$U_X = \sum_{\text{только для}\{x_i\}} r_{qx} \ ,$$

$$U_Y = \sum_{\text{только для}\{y_i\}} r_{qy} \ ,$$

в которых просуммированы только номера из общего вариационного ряда, соответствующие только одной выборке. Если нулевая гипотеза верна, то при $n_1+n_2 \geq 20$ и $min(n_1, n_2) \geq 5$ распределение критерия Вилкоксона близко к нормальному со следующими математическими ожиданиями и дисперсией:

$$M_{U_x} = \frac{1}{2} n_1 (n_1+n_2+1) \ ,$$

$$M_{U_y} = \frac{1}{2} n_2 (n_1+n_2+1) \ ,$$

$$\sigma_U^2 = \frac{1}{12} n_1 n_2 (n_1+n_2+1) \ .$$

Для проверки нулевой гипотезы определяются границы критической области для показателей

$$U_{HX} = M_{U_x} - u_\gamma \sigma_U; \quad U_{BX} = M_{U_x} + u_\gamma \sigma_U \ ;$$

$$U_{HY} = M_{U_y} - u_\gamma \sigma_U; \quad U_{BY} = M_{U_y} + u_\gamma \sigma_U \ ,$$

где $u_\gamma$ это квантиль стандартного нормального распределения для $\gamma = 1-\frac{\alpha}{2}$ (в R $u_\gamma = qnorm(\gamma, 0, 1)$ ). Решающее правило формулируется следующим образом:

гипотеза $H_0: f(x) = f(y)$ принимается, если $U_X \in [U_{HX}; U_{BX}]$ и $U_Y \in [U_{HY}; U_{BY}]$ ;

иначе гипотеза $H_0$ отклоняется.

## Проверка гипотез о типе распределения

При априорной неизвестности типа распределения СВ, он наиболее достоверно определяется на основе гистограммы распределения по $\chi^2$ -критерию К.Пирсона или на основе эмпирической функции распределения по $\omega^2$ критерию Крамера-Мизеса-Смирнова.

### $\chi^2$ -критерий

Пусть имеется выборка $x_i, i=1,2,\ldots,n$ и предполагается распределение СВ в виде $f(x,\boldsymbol{\theta})$ , где $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_J)$ . Строится гистограмма из *K* отрезков длиной $\Delta$ , определяются числа $n_k$ элементов выборки попавших в *k*-ый отрезок. Теоретическая вероятность попадания в *k*-ый отрезок с выбранной функцией плотности вероятности

$$p_{kT} = f(x_k,\boldsymbol{\theta})\cdot\Delta, \quad k=1,2,\ldots,K ,$$

где $x_k$ - средняя точка *k*-ого отрезка гистограммы. Параметры $\boldsymbol{\theta}$ оцениваются на основе статистических методов. С другой стороны, на основе выборки легко определяется экспериментальная вероятностью

$$p_{k\,эксп} = \frac{n_k}{n}, \quad k=1,2,\ldots,K .$$

Используется критерий

$$\chi^2 = n\sum_{k=1}^{K}\frac{(p_{kT}-p_{k\,эксп})^2}{p_{kT}} ,$$

который при нормальном распределении СВ имеет точно $\chi^2$ -распределение с $q=K-J-1$ степенями свободы. Доказано, что в пределе, при $n\to\infty$ распределение показателя стремиться к $\chi^2$ -распределению и при других распределениях СВ. Проверка гипотезы осуществляется следующим образом. Выбирается уровень значимости $\alpha$ и определяется критический показатель из условия $P\{\chi^2\ge\chi^2_{кр}\}=\alpha$ , что в R можно сделать через вызов функции `qchisq()`, т. е.

$$\chi^2_{кр} = qchisq(1-\alpha, q) .$$

Затем вычисляется реализация $\chi^2_{эксп}$ и сравнивается с критическим показателем, т. е. решающее правило имеет вид:

гипотеза $H_0: f(x)=f(x,\boldsymbol{\theta})$ принимается, если $\chi^2_{эксп} < \chi^2_{кр}$ ;

гипотеза $H_0$ отклоняется, если $\chi^2_{эксп} \geq \chi^2_{кр}$ .

## $\omega^2$ -критерий Крамера-Мизеса-Смирнова

Данный критерий основан на непосредственном сравнении функции распределения $F(x,\theta)$ с эмпирической ФР, основанной на вариационном ряде. В качестве меры отклонения гипотетической ФР от эмпирической используется средний квадрат разности функций распределения по всем значениям аргумента:

$$\omega^2 = n\int_{-\infty}^{\infty}\left[F_Э(x) - F(x)\right]^2 dF(x) \ .$$

После преобразования интеграла к сумме, со значениями в точках выборки, с учетом зависимости эмпирической ФР от упорядоченных элементов выборки, показатель принимает вид

$$\omega^2 = s = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(x_i) - \frac{2i-1}{2n}\right]^2 \ .$$

Математическое ожидание и дисперсия этого показателя

$$M_{\omega^2} = \frac{1}{6}\ ; \qquad \sigma^2_{\omega^2} = \frac{4n-3}{180n} \ .$$

Известно, что при $n > 40$ распределение этого показателя близко к предельному, которое можно использовать для вычисления квантилей (хотя можно найти таблицы). Предельное распределение имеет вид

$$a1(s) = \frac{1}{\sqrt{2s}}\sum_{j=0}^{\infty}\frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2}{16s}\right\}\left\{I_{-\frac{1}{4}}\left(\frac{(4j+1)^2}{16s}\right) - I_{\frac{1}{4}}\left(\frac{(4j+1)^2}{16s}\right)\right\} \ ,$$

где $I_{-\frac{1}{4}}$ и $I_{\frac{1}{4}}$ - модифицированные функции Бесселя, которые можно представить в виде следующего выражения

$$I_\nu(z) = \sum_{k=0}^{\infty}\frac{\left(\frac{z}{2}\right)^{\nu+2k}}{\Gamma(k+1)\Gamma(k+\nu+1)} \ , \qquad |z| < \infty, \qquad |arg\, z| < \pi \ .$$