Статистика

Алексей Померанцев

Содержание

1 Введение

В этом документе собраны основные сведения из математической статистики, которые используются в хемометрике. Приведенный текст не может служить учебником по статистике – это скорее конспект, краткий справочник по математической статистике. Более глубокое и систематическое изложение может быть найдено в литературе.

Изложение иллюстрируется примерами, выполненными в рабочей книге Excel, Statistics.xls, которая сопровождает этот документ.

2 Базовые сведения

2.1 Вероятность события

В мире часто происходят события, исход которых не предопределен заранее. Всем известен хрестоматийный пример с подбрасыванием монетки, завершающийся случайным событием: выпадением орла или решки. Таким случайным событиям можно приписать вероятность – число от нуля до единицы. Однако не у всякого события может быть вероятность. Ключевым условием является повторяемость. Поэтому бессмысленно спрашивать, какова вероятность того, что завтра пойдет дождь. У «завтра» нет повторяемости – это уникальное событие, которое нельзя повторить. Однако можно говорить о вероятности того, что 7 июля будет дождь. Событие «7 июля» повторяется каждый год, и дождю в этот день можно приписать некоторую вероятность.

Понятие вероятности можно применять только к тем событиям, которые еще не произошли, или исход которых нам пока не известен. Так, например, мы можем рассчитать вероятность выигрыша в лотерею, но, как только нам стал известен результат розыгрыша, т.е. событие уже произошло – рассчитанная вероятность теряет всякий смысл.

Еще одним важным понятием является пространство событий – это полный набор всех возможных исходов. Так в опыте с монеткой есть только два события: орел и решка. Рассмотрим другой опыт – измерение роста случайно выбранного человека. Если точность измерения один сантиметр, то пространство событий – это набор чисел от 30 см (новорожденный), до 251 см (рекорд книги Гиннеса) – всего 222 варианта. Однако если мы меряем рост с точность до 1 метра, то в пространстве оказываются только три события: меньше 1 м, от 1 м до 2 м, и больше 2 м.

2.2 Случайная величина

Случайная величина — это переменная, значение которой до опыта (реализации) неизвестно. Всякая случайная величина характеризуется:

- множеством своих возможных значений (пространство событий)
- неограниченным числом повторения реализаций
- вероятностью попадания в любую наперед заданную область во множестве значений

Множество значений может быть дискретным, непрерывным и дискретно-непрерывным. Соответственно именуются и случайные величины.

2.3 Распределение случайной величины

Пусть X – это случайная величина, множеством возможных значений которой являются действительные числа. Рассмотрим вероятность события, что реализация X не больше заданного числа x. Если рассматривать эту вероятность в зависимости от величины x, то получится функция $F(x)$, называемая (кумулятивной) *функцией распределения* случайной величины –

$$F(x) = \Pr\{X \leq x\}$$

Функция распределения это неубывающая функция, которая стремится к 0 при малых x, и стремится к 1 при больших значениях аргумента.

То, что случайная величина X имеет функцию распределения F обозначается так –

$$X \sim F$$

Распределение называется симметричным (относительно точки a) если $F(a + x) = 1 - F(a - x)$.

Для дискретных случайная величина функция распределения кусочно-постоянна со скачками в точках $x = x_i$.

Производная функция распределения $F(x)$ называется *плотностью вероятности* $f(x)$:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

2.4 Математическое ожидание

Пусть X – это случайная величина с плотностью вероятности $f(x)$.

Математическим ожиданием X называется величина

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x f(x)dx$$

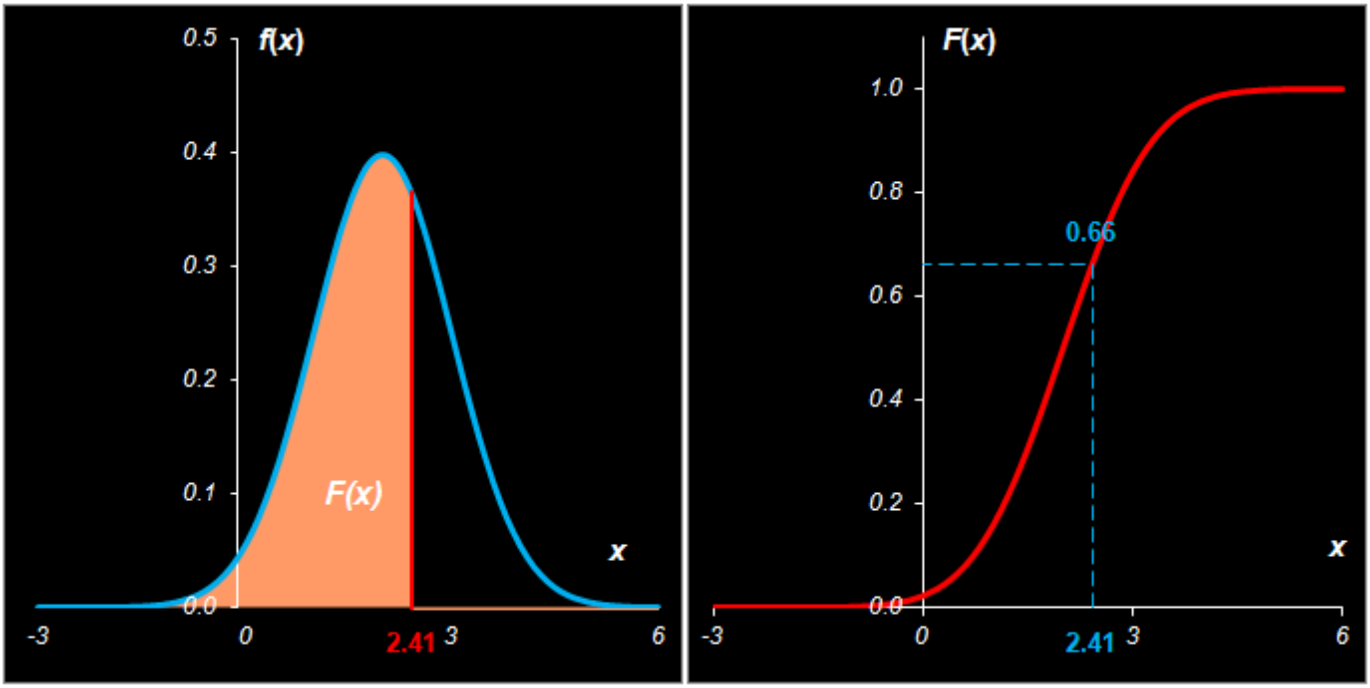


Рис. 2.1. Плотность вероятности $f(x)$ и функция распределения $F(x)$ случайной величины

2.5 Дисперсия

Пусть X – это случайная величина с плотностью вероятности $f(x)$.

Дисперсией X называется величина

$$\mathrm{V}(X) = \int_{-\infty}^{+\infty} \big(x - \mathrm{E}(X)\big)^2 f(x)dx = \mathrm{E}\big(X - \mathrm{E}(X)\big)^2$$

Если из дисперсии извлечь квадратный корень, то получится величина, называемая *среднеквадратичным отклонением* (СКО).

2.6 Моменты

Пусть X – это случайная величина с плотностью вероятности $f(x)$.

Моментом порядка n называется величина

$$\mu_n = \int_{-\infty}^{+\infty} x^n f(x)dx$$

По определению $\mu_1 = \mathrm{E}(X)$.

Центральным моментом порядка n называется величина

$$m_n = \int_{-\infty}^{+\infty} (x - \mu_1)^n f(x) dx = \mathrm{E}\big((X - \mu_1)^n\big)$$

По определению $m_2 = \mathrm{V}(X)$.

2.7 Квантили

Пусть $F(x)$ – (кумулятивная) функция распределения случайной величины

$$F(x) = \int_{-\infty}^{+\infty} f(t) dt$$

Рассмотрим функцию $F^{-1}(P), 0 \le P \le 1$, обратную к $F(x)$ т.е. $F^{-1}(F(x)) = x$ и $F(F^{-1}(P)) = P$. Функция $F^{-1}(P)$ называется *P-квантилем* распределения F.

Величина квантиля для $P = 0.5$ называется *медианой* распределения.

Квантили для $P = 0.25, 0.75$ называются *квартилями*, а для $P = 0.01, 0.02, ..., 0.99$ называются *процентилями*.

2.8 Многомерные распределения

Две (и более) случайные величины можно рассматривать совместно. Совместная (кумулятивная) функция распределения двух случайных величин X и Y определяется так

$$F(x, y) = Pr\big\{(X \le x) \wedge (Y \le y)\big\} = \int_{-\infty}^{x} \int_{-\infty}^{y} f(\xi, \eta) d\xi d\eta$$

Так же, как и в одномерном случае, функция $f(x, y)$ называется плотностью вероятности .

Случайные величины X и Y называются *независимыми*, если их совместная плотность вероятности равна произведению частных плотностей.

$$f(x, y) = f(x) f(y)$$

2.9 Ковариация и корреляция

Ковариацией случайных величин X и Y называется (детерминированная) величина

$$\mathrm{cov}(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\xi - \mathrm{E}(X))(\eta - \mathrm{E}(Y)) f(\xi, \eta) d\xi d\eta$$

где $f(x, y)$ – совместная плотность вероятности. Величина

$$\mathrm{cor}(X,Y) = \frac{cov(X,Y)}{\sqrt{\mathrm{V}(X)\mathrm{V}(Y)}}$$

называется *корреляцией* случайных величин X и Y.

Если случайные величины X и Y независимы, то их ковариация и корреляция равны нулю. Обратное не верно.

Для совместных распределений многомерных случайных величин $X_1, \dots, X_n$ *ковариационная матрица* **C**

$$c_{ij} = cov(X_i, X_j), i, j = 1, \dots, n$$

играет ту же роль, что и дисперсия в одномерном распределении.

2.10 Функции от случайной величины

Функция от случайной величины также является случайной величиной.

Пусть случайная величина X имеет функцию распределения $F_X(x)$, и случайные величины X и Y связаны взаимно однозначными соотношениями $y = \varphi(x)$, $x = \psi(y)$.

Если $\varphi(x)$ – возрастающая функция, то функция распределения и квантили случайной величины Y определяются так:

$$F_Y(y) = F_X(\psi(y))$$

$$y(P) = \psi(x(P))$$

Если $\psi(y)$ – дифференцируемая функция, то плотность вероятности случайной величины Y вычисляется по формуле:

$$f_Y(y) = f_X(\psi(y))\left|\frac{d\psi}{dy}\right|$$

Для линейных преобразований $y = ax + b$

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

Кроме этого:

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$$

$$\mathrm{V}(aX + b) = a^2\mathrm{V}(X)$$

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

$$\mathrm{V}(X + Y) = \mathrm{V}(X) + \mathrm{V}(Y) + \mathrm{cov}(X, Y)$$

2.11 Стандартизация случайной величины

Если случайная величина X имеет математическое ожидание m и дисперсию s^2: $\mathrm{E}(X) = m$, $\mathrm{V}(X) = s^2$, то случайная величина:

$$Y = (X - m)/s$$

называется *стандартизованной* (нормированной), поскольку $\mathrm{E}(Y) = 0$, $\mathrm{V}(Y) = 1$.

3 Основные распределения

3.1 Биномиальное распределение

Дискретная случайная величина X имеет дискретное биномиальное распределение, если ее плотность вероятности имеет вид

$$f(k|p,n) \equiv \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

где $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ – биномиальный коэффициент.

Биномиальное распределение – это распределение числа успехов k в серии из независимых n опытов, при условии, что вероятность успеха в каждом опыте есть p.

Математическое ожидание и дисперсия, соответственно, равны

$$\mathrm{E}(X) = np$$

$$\mathrm{V}(X) = np(1-p)$$

При больших n биномиальное распределение хорошо приближается нормальным.

Для вычисления биномиального распределения в Excel используется стандартная функция `BINOMDIST` (`БИНОМРАСП`):

```
BINOMDIST(number_s=k, trials=n, probability_s=p,cumulative=TRUE|FALSE)
```

Если `cumulative=TRUE`, то возвращается кумулятивная функция распределения, а если `cumulative=FALSE`, то возвращается плотность вероятности.

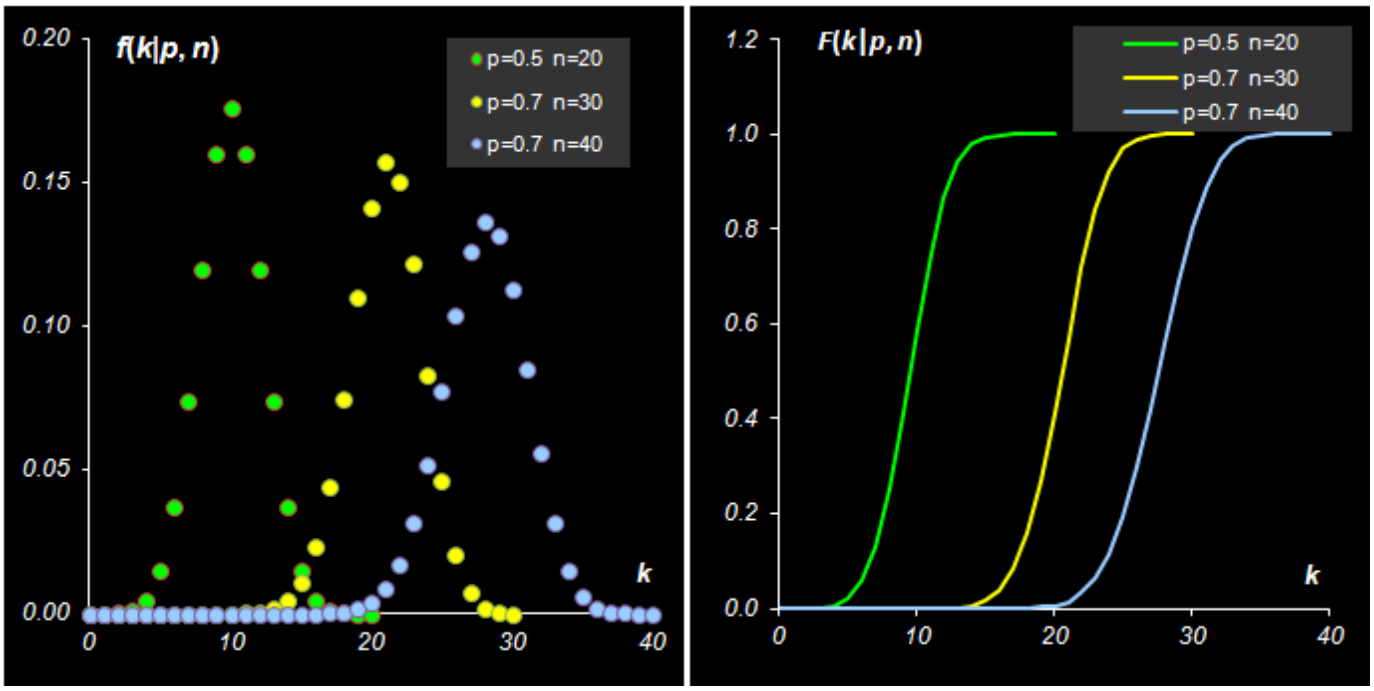


Рис. 3.1. Плотность вероятности и функция распределения биномиального распределения

	A	B	C	D	E	F	G	H	I
2		*n*=	20.000	30.000	40.000				
3		*p*=	0.500	0.700	0.700				
4			Density *f*(*k*|p,n)				Cumulative *F*(k|p,n)		
5	*k*	0	0.000	0.000	0.000		0.000	0.000	0.000
6		1	0.000	=BINOMDIST($B6,D$2,D$3,FALSE)					0.000
7		2	0.000	0.000	0.000		0.000	0.000	0.000
8		3	0.001	0.000	0.000		0.001	0.000	0.000
9		4	0.005	0.000	0.000		0.006	0.000	0.000
10		5	0.015	0.000	0.000		0.021	0.000	0.000
11		6	0.037	0.000	0.000		0.058	0.000	0.000

Рис. 3.2. Пример вычисления биномиального распределения

3.2 Равномерное распределение

Случайная величина X распределена равномерно на отрезке $[a, b]$, если ее функция распределения $U(x|a, b)$ и, соответственно, плотность вероятности $u(x|a, b)$ имеют вид

$$U(x|a,b) = \begin{cases} 0, x \le a, \\ \frac{x-a}{b-a}, a < x \le b \\ 1, x > b \end{cases}$$

$$u(x|a,b) = \begin{cases} 0, x \le a, \\ \frac{1}{b-a}, a < x \le b \\ 0, x > b \end{cases}$$

Математическое ожидание и дисперсия, соответственно, равны

$$\mathrm{E}(X) = 0.5(a+b)$$

$$\mathrm{V}(X) = (b-a)^2/12$$

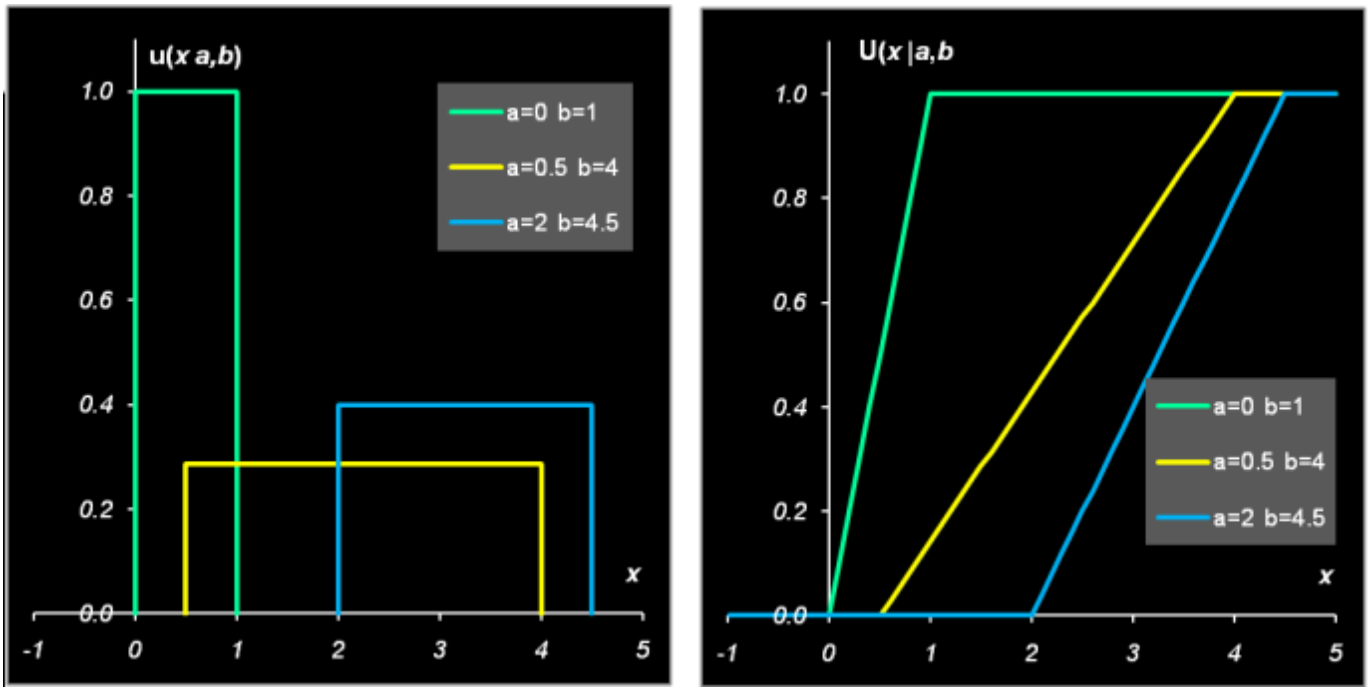


Рис. 3.3. Плотность вероятности и функция распределения равномерного распределения

То, что случайная величина X распределена равномерно на отрезке $[a, b]$, будем обозначать:

$$X \sim U(a, b)$$

3.3 Нормальное распределение

Нормальное (или гауссово) распределение – это, наверное, самое важное распределение в статистике. Плотность этого распределения имеет вид

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

Нормальное распределение зависит от двух параметров: μ и σ^2 и оно обычно обозначается $N(\mu, \sigma^2)$ т.е.

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

Математическое ожидание и дисперсия нормального распределения равны, соответственно:

$$\mathrm{E}(X) = \mu, \mathrm{V}(X) = \sigma^2$$

Нормальное распределение называется *стандартным*, если $\mu = 0$, $\sigma^2 = 1$.

Если $X_0 \sim N(0, 1)$, то $X = \mu + \sigma X_0 \sim N(\mu, \sigma^2)$.

Кумулятивная функция стандартного нормального распределения:

$$\Phi(x) = \int_{-\infty}^{x} f(t)dt$$

является специальной функцией, т.к. она не выражается через элементарные функции.

Квантили стандартного нормального распределения обозначаются $\Phi^{-1}(P)$.

Стандартное нормальное распределение симметрично, поэтому для него верны следующие соотношения:

$$\Phi(-x) = 1-\Phi(x)$$

$$\Phi^{-1}(1-P) = -\Phi^{-1}(P)$$

Для вычисления нормального распределения в Excel используется стандартные функции: `NORMDIST` (`НОРМРАСП`) и `NORMSDIST` (`НОРМСТРАСП`), а также `NORMINV` (`НОРМОБР`) и `NORMSINV` (`НОРМСТОБР`).

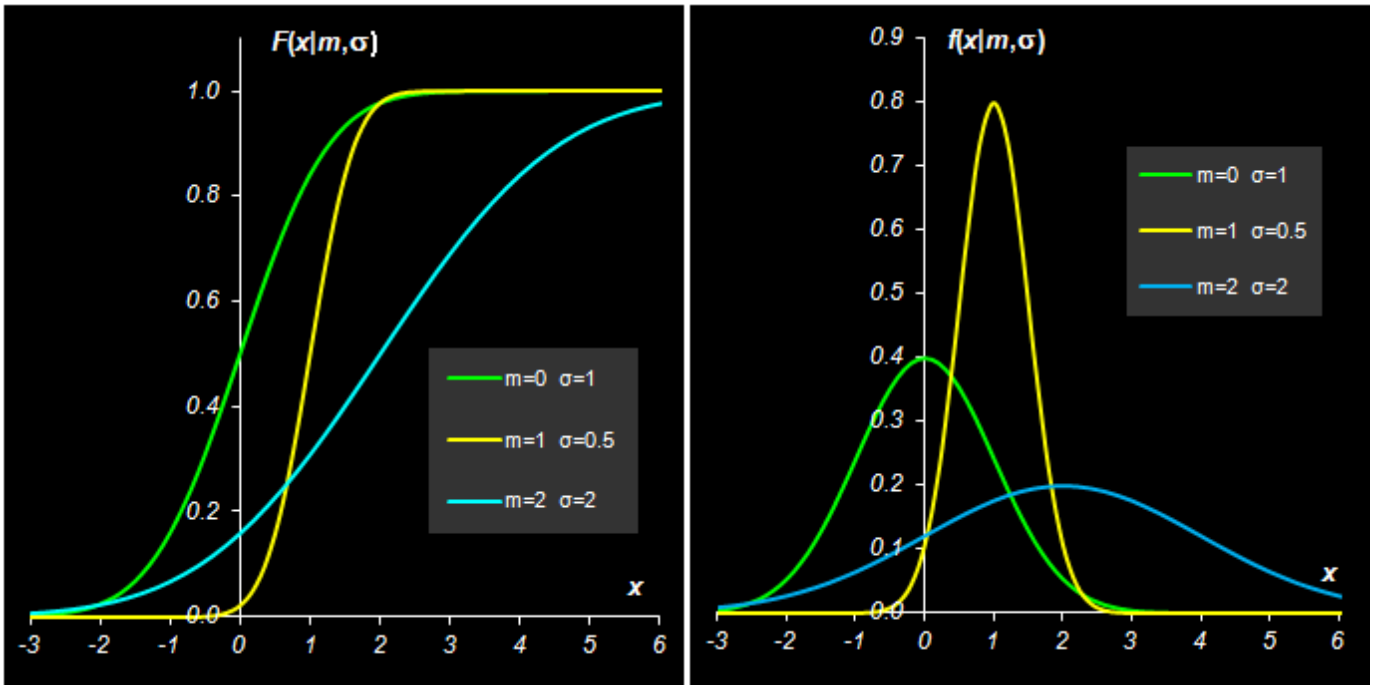


Рис. 3.4. Функция распределения и плотность вероятности нормального распределения

Синтаксис:

```
NORMDIST(x, mean=mu, standard_dev=sigma, cumulative=TRUE|FALSE)
```

Если `cumulative=TRUE` то возвращается кумулятивная функция распределения $\Phi(x|\mu,\sigma^2)$, а если `cumulative=FALSE`, то возвращается плотность вероятности, $f(x|\mu,\sigma^2)$.

```
NORMSDIST(x)
```

Возвращается кумулятивная функция стандартного нормального распределения в точке x.

```
NORMINV(probability=P, mean=mu, standard_dev=sigma)
```

Возвращается квантиль $\Phi^{-1}(P|\mu,\sigma^2)$ нормального распределения для вероятности P.

```
NORMSINV(probability=P)
```

Возвращается квантиль $\Phi^{-1}(P|0,1)$ стандартного нормального распределения для вероятности P.

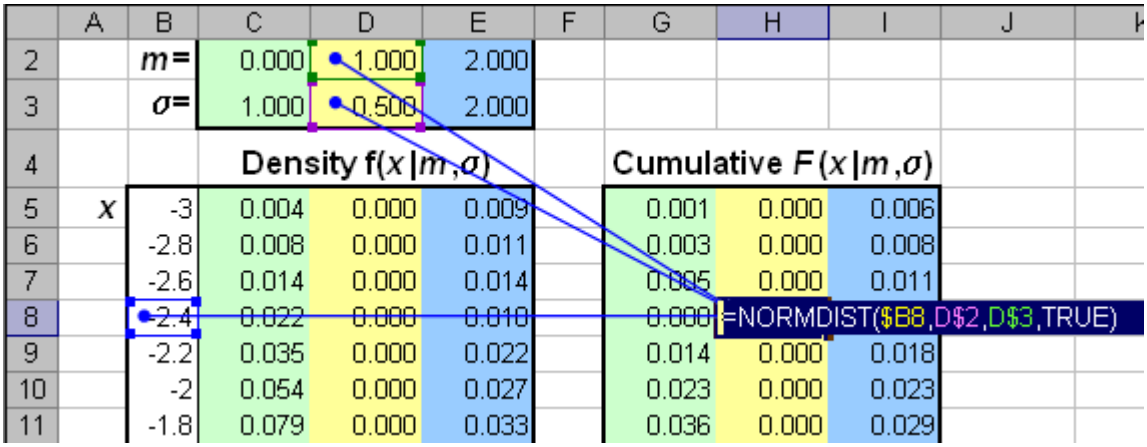


Рис. 3.5. Пример вычисления нормального распределения

3.4 Распределение хи-квадрат

Рассмотрим N независимых стандартных нормальных случайных величин $X_1, \ldots, X_N$ с нулевым мат. ожиданием и единичной дисперсией, т.е.

$$X_n \sim N(0,1)$$

Величина

$$\chi^2(N) = X_1^2 + X_2^2 + \cdots + X_N^2$$

является случайной, распределение которой носит название хи-квадрат. Это распределение зависит от одного параметра – N, который называется *числом степеней свободы.* Плотность вероятности распределения хи-квадрат имеет вид

$$f(x|N) = \frac{(1/2)^{\frac{N}{2}}}{\Gamma\left(\frac{N}{2}\right)} x^{\frac{N}{2}-1} e^{-\frac{x}{2}}$$

Распределение хи-квадрат широко используется в статистике, например, при проверке гипотез.

Математическое ожидание и дисперсия распределения χ2(N) равны, соответственно,

$$\mathrm{E}(\chi^2(N)) = N$$

$$\mathrm{V}(\chi^2(N)) = 2N$$

При больших N распределение хи-квадрат хорошо приближается нормальным с этими же параметрами.

Квантили распределения $\chi^2(N)$ обозначаются $\chi^{-2}(P|N)$.

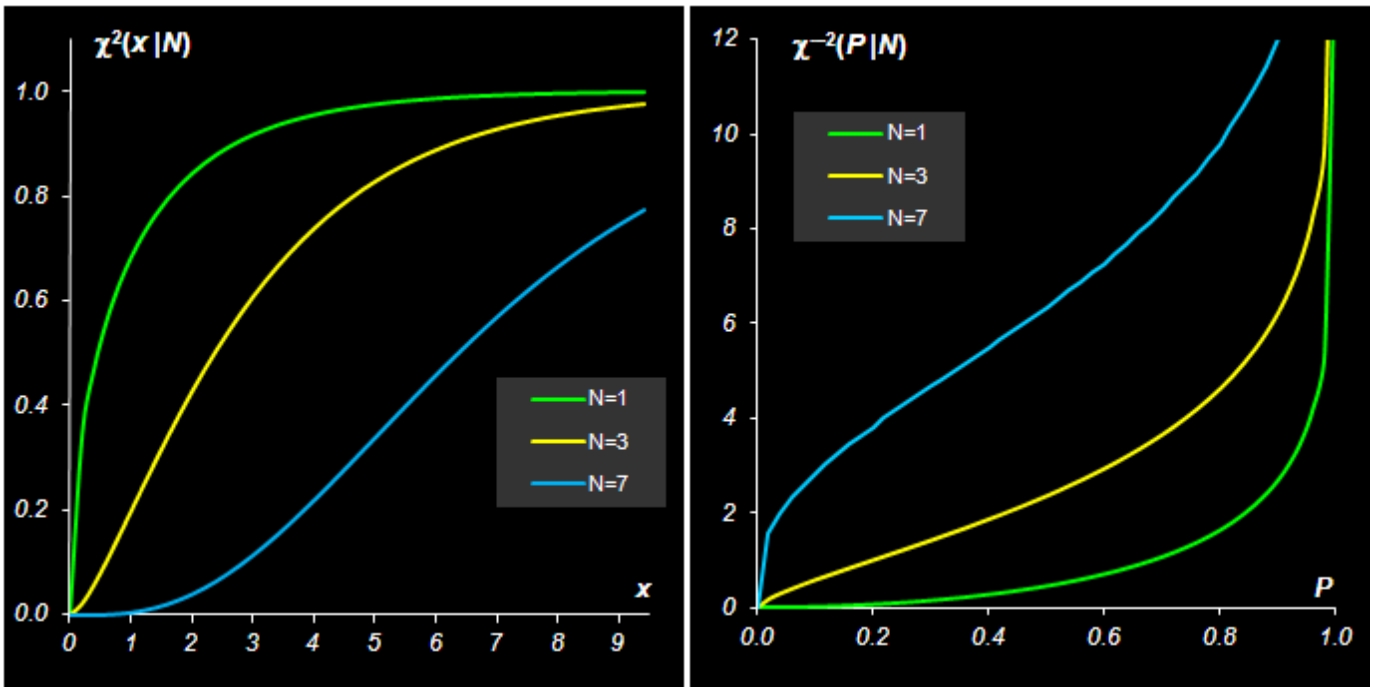


Рис. 3.6. Функция распределения и квантиль распределения хи-квадрат

Для вычисления распределения хи-квадрат в Excel используется две стандартные функции: `CHIDIST` (`ХИ2РАСП`) и `CHIINV` (`ХИ2ОБР`).

```
CHIDIST(x, degrees_freedom=N)
```

Возвращается значение $1-\chi^2(x|N)$, где $\chi^2(x|N)$ – кумулятивная функция распределения хи-квадрат.

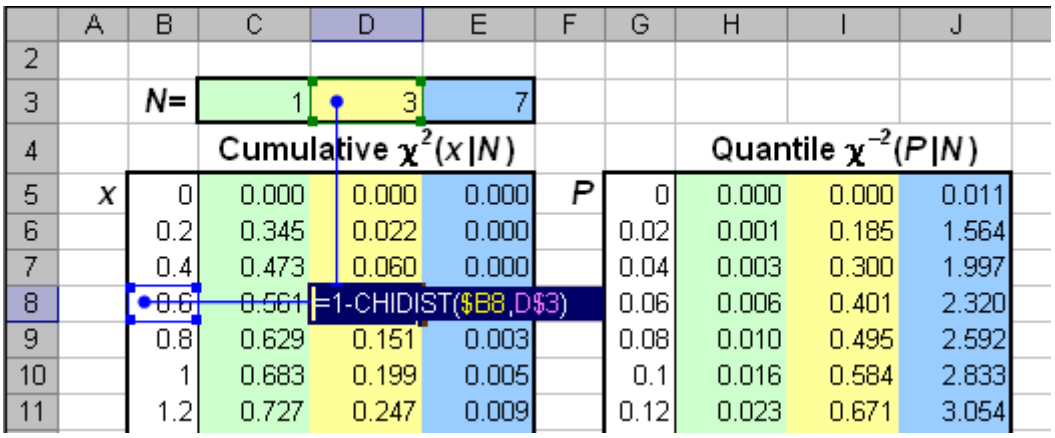
```
CHIINV(probability=1–P ,degrees_freedom=N)
```

Возвращается квантиль $\chi^{-2}(1-P|N)$ распределения хи-квадрат для вероятности $1-P$.

3.5 Распределение Стьюдента

Рассмотрим две случайные величины: X – распределенную стандартно-нормально $X \sim N(0,1)$, и Y – распределенную по хи-квадрат с N степенями свободы $Y \sim \chi^2(N)$.

Случайная величина



	x		Cumulative $\chi^2(x\|N)$			P		Quantile $\chi^{-2}(P\|N)$		
		N=	1	3	7			1	3	7
5		0	0.000	0.000	0.000		0	0.000	0.000	0.011
6		0.2	0.345	0.022	0.000		0.02	0.001	0.185	1.564
7		0.4	0.473	0.060	0.000		0.04	0.003	0.300	1.997
8		0.6	0.561	=1-CHIDIST($B8,D$3)			0.06	0.006	0.401	2.320
9		0.8	0.629	0.151	0.003		0.08	0.010	0.495	2.592
10		1	0.683	0.199	0.005		0.1	0.016	0.584	2.833
11		1.2	0.727	0.247	0.009		0.12	0.023	0.671	3.054

Рис. 3.7. Пример вычисления распределения хи-квадрат

$$T(N) = \sqrt{N}\frac{X}{\sqrt{Y}}$$

подчиняется распределению, которое носит имя Стьюдента. Это распределение зависит от одного параметра N, который также называется числом степеней свободы. Распределение Стьюдента применяется в проверке гипотез и для построения доверительных интервалов.

Математическое ожидание $T(N)$ равно нулю, а дисперсия равна

$$\mathrm{V}(T(N)) = N/(N-2), N > 2$$

Распределение Стьюдента симметрично, и при $N > 20$ неотличимо от нормального.

Формула для плотности вероятности Стьюдента приведена во многих пособиях. Квантили распределения $T(N)$ обозначаются $T^{-1}(P|N)$.

Для вычисления распределения Стьюдента в Excel используется две стандартные функции: `TDIST` (`СТЬЮДРАСП`) и `TINV` (`СТЬЮДРАСПОБР`).

```
TDIST(x, degrees_freedom=N, tails=1|2)
```

Если `tails=1`, то функция `TDIST` возвращает значение Pr
$T(N) > x$, а при `tails=2` значение Pr
$|T(N)| > x$. Значения при $x < 0$ не возвращаются. Поэтому, для того, чтобы вычислить в Excel обычную кумулятивную функцию распределения Стьюдента $T(x|N)$, приходится использовать следующую формулу

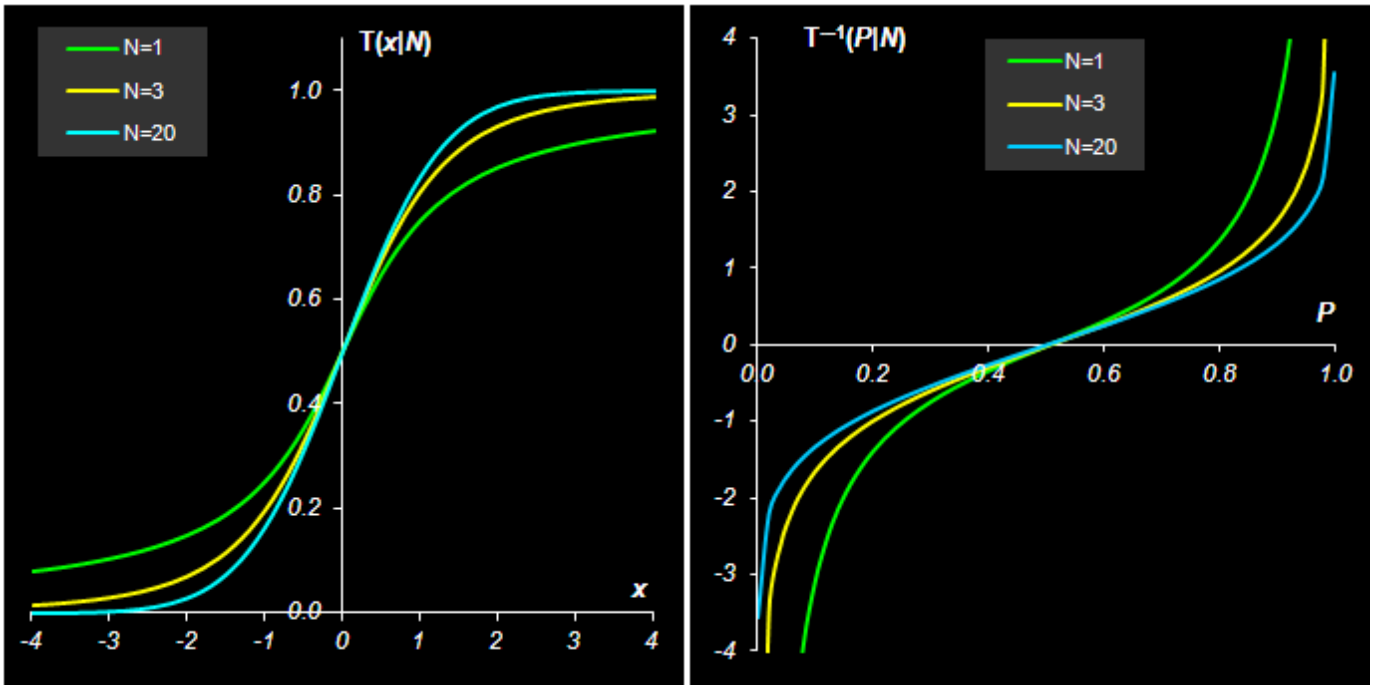


Рис. 3.8. Функция распределения и квантиль распределения Стьюдента

```
IF(x>0, 1-TDIST(x,N,1), -TDIST(-x,N,1))
```

Функция:

```
TINV(P, degrees_freedom=N)
```

возвращает значение x, для которого Pr
$|T(N)| > x = P$. И в этом случае для вычисления в Excel квантиля распределения Стьюдента $T^{-1}(P|N)$, нужно использовать следующую формулу

```
IF(P<0.5, TINV(2*P,N), -TINV(2-2*P,N)).
```

3.6 Распределение Фишера

Пусть имеются две независимые случайные величины X_1 и X_2, каждая из которых подчиняется распределению хи-квадрат с N_1 и N_2 степенями свободы, т.е.

$$X_1 \sim \chi^2(N_1)$$

$$X_2 \sim \chi^2(N_2)$$

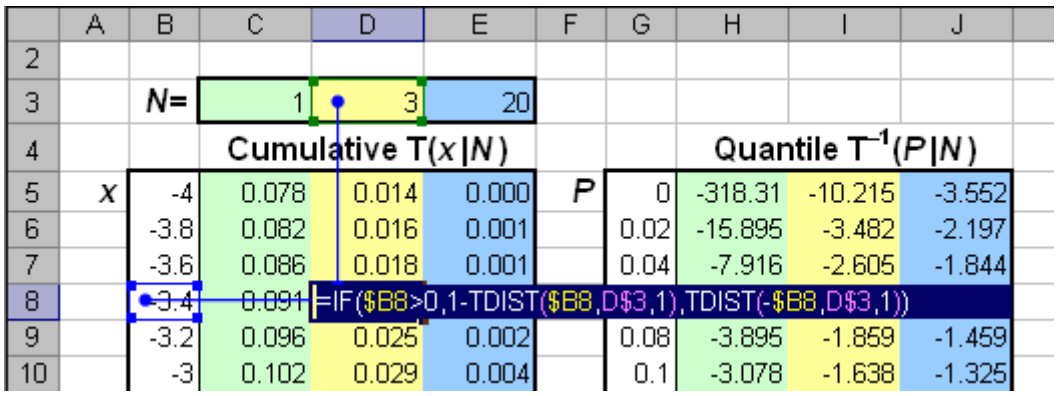


Рис. 3.9. Пример вычисления распределения Стьюдента

Случайная величина:

$$F(N_1, N_2) = \frac{N_2 X_1}{N_1 X_2}$$

подчиняется распределению, которое носит имя Фишера. Это распределение зависит от двух параметров N_1 и N_2, которые также называются числами степеней свободы. Математическое ожидание и дисперсия распределения $F(N_1, N_2)$ равны, соответственно:

$$\mathrm{E}(F(N_1, N_2)) = N_2/(N_2-2), N_2 > 2$$

$$\mathrm{V}(F(N_1, N_2)) = \frac{2N_2^2(N_1 + N_2 - 2)}{N_1(N_2-2)^2(N_2-4)}, N_2 > 4$$

Формула для плотности вероятности распределения Фишера приведена во многих пособиях.

Если $X \sim F(N_1, N_2)$, то $1/X \sim F(N_2, N_1)$.

Квантили распределения $F(N_1, N_2)$ обозначаются $F^{-1}(P|N_1, N_2)$.

Для вычисления распределения Фишера в Excel используются две стандартные функции: `FDIST` (`FРАСП`) и `FINV` (`FРАСПОБР`).

```
FDIST(x, degrees_freedom1=N1, degrees_freedom2=N2)
```

Возвращается значение $1-F(x|N1, N2)$, где $F(x|N1, N2)$ – кумулятивная функция распределения Фишера.

```
FINV(probability=1–P, degrees_freedom1=N1, degrees_freedom2=N2)
```

Возвращается квантиль $F^{-1}(1-P|N1, N2)$ для вероятности $1-P$.

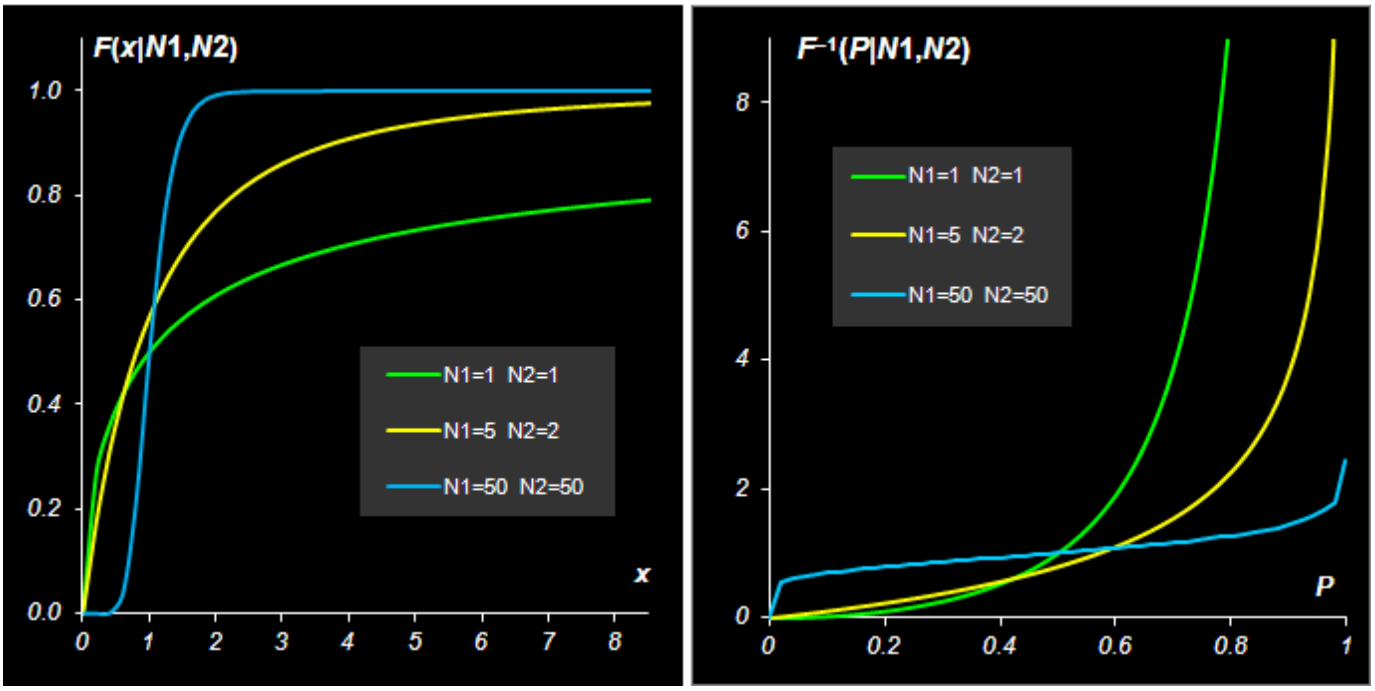


Рис. 3.10. Функция распределения и квантиль распределения Фишера

	A	B	C	D	E	F	G	H	I	J	K
2		*N* 1=	1	2	50						
3		*N* 2=	1	5	50						
4			Cumulative F(*x*|*N1*,*N2*					Quantile F^{-1}(*P*|*N1*,*N2*)			
5	*x*	0	0.000	0.000	0.000	*P*	0	0.000	0.000	0.000	
6		0.2	0.268	0.175	0.000		0.02	0.001	0.020	0.555	
7		0.4	0.359	0.310	0.001		0.04	0.004	0.041	0.606	
8		0.6	0.420	0.416	0.037		0.06	0.009	0.063	0.642	
9		0.8	0.465	0.500	0.216		0.08	0.016	=FINV(1-$G9,D$2,D$3)		
10		1	0.500	0.569	0.500		0.1	0.025	0.108	0.694	
11		1.2	0.529	0.625	0.739		0.12	0.036	0.131	0.715	
12		1.4	0.553	0.671	0.881		0.14	0.050	0.155	0.735	

Рис. 3.11. Пример вычисления распределения Фишера

3.7 Многомерное нормальное распределение

Это распределение является естественным обобщением одномерного нормального распределения на случай многомерной случайной величины, т.е. случайного вектора $\mathbf{x}$, размерностью n.

Функция плотности вероятности имеет следующий вид

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\det(\boldsymbol{\Sigma})(2\pi)^n} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{t}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

где $\boldsymbol{\Sigma}$ – симметричная положительно определенная $(n \times n)$ матрица.

Многомерное нормальное распределение зависит от двух групп параметров:

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Математическое ожидание $\mathbf{x}$ равно $\boldsymbol{\mu}$, а ковариационная матрица равна матрице $\boldsymbol{\Sigma}$.

3.8 Генерация случайных чисел

Иногда бывает полезно создать искусственную выборку случайных чисел, подчиняющихся заданному распределению. Это можно сделать, используя следующее простое утверждение.

Пусть $F(x)$ и $F^{-1}(P)$ суть некоторая функция распределения и ее квантиль, соответственно. Если случайная величина X распределена равномерно на отрезке $[0, 1]$, т.е

$$X \sim U(0, 1)$$

тогда случайная величина

$$Y = F^{-1}(X)$$

имеет функцию распределения F.

Таким образом, если получить набор случайных величин, распределенных равномерно, то эти случайные величины можно превратить в новые, имеющие другое, заданное распределение.

Для генерации случайных чисел в Excel имеется стандартная функция: `RAND` (`СЛЧИС`).

```
RAND()
```

Возвращает случайное число, равномерно распределенное на отрезке $[0,1]$. Новое случайное число возвращается при каждом вычислении рабочего листа.

На листе *Random* рабочей книги *Statistics.xls* приведен пример генерации случайных чисел для разных распределений.

	A	B	C	D	E	F	G
2		*m*=	0		*N* 1=	5	
3		*s*=	1		*N* 2=	10	
4							
5	n	U(0,1)	N(0,1)	χ^2(N1)	T(N1)	F(N1,N2)	
6	1	0.849	1.033	8.099	1.151	2.086	
7	2	0.315	-0.482	3.095	-0.514	0.625	
8	3	0.751	0.679	6.642	0.732	1.591	
9	4	0.792	0.814	=CHIINV(1-$B9,N_1)		1.765	
10	5	0.130	-1.128	1.843	-1.272	0.351	
11	6	0.218	-0.779	2.463	-0.847	0.482	
12	7	0.749	0.673	6.619	0.725	1.583	

Рис. 3.12. Пример генерации случайных чисел

4 Оценка параметров

4.1 Выборка

Предположим, что имеется набор чисел $\mathbf{x} = (x_1, \ldots, x_I)$, и каждое x_i является одной реализацией случайной величины, подчиняющейся, вообще говоря, неизвестному распределению. Этот набор называется *выборкой*, а число I – объемом выборки.

В случае одномерного распределения выборка – это вектор $\mathbf{x}$, а в многомерном случае выборка – это матрица $\mathbf{X}$ размерностью $I \times J$, каждая строка которой представляет одну реализацию (наблюдение) многомерной случайной величины размерностью J.

Обычно предполагается, что все элементы выборки статистически независимы. В практических приложениях слово «выборка» часто заменяется словом «данные».

4.2 Выбросы и маргиналы

Среди элементов выборки могут присутствовать такие, которые существенно отличаются от других элементов.

Пусть, например, имеется выборка из стандартного нормального распределения $N(0, 1)$, в которой присутствует элемент со значением $x_{out} = 3.2$. Для такого распределения вероятность единичного события $x_{out} \geq 3.2$ мала – она равна $\alpha = 0.0007$. Однако значение x_{out} присутствует в независимой выборке размера I, поэтому нужно рассчитывать вероятность события «хотя бы один раз среди I попыток»

$$P_{out} = 1 - (1 - \alpha)^I \approx 1 - \exp(-I\alpha)$$

Для $I = 10$ $P_{out} = 0.007$, для $I = 100$ $P_{out} = 0.07$, а для $I = 1000$ $P_{out} = 0.50$. Естественно – чем больше выборка, тем выше вероятность того, что встретится такое экстремальное значение.

Таким образом, интерпретация выпадающих из выборки значений существенно зависит от объема выборки – для малых I их нужно рассматривать как *выбросы* (промахи при измерениях) и, соответственно, удалять из выборки. Для больших I такие выпадающие значения являются приемлемыми *маргиналами* и они должны сохраняться в выборке.

4.3 Генеральная совокупность

Операцию создания выборки в статистике называют *извлечением*. Тем самым подчеркивают, что имеющаяся у нас выборка $\mathbf{x}_1$ не единственная, и что можно получить (часто только теоретически) и другие похожие выборки $\mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n$. Слово *похожие* означает, что все эти выборки устроены аналогичным способом – подчиняются одному и тому же распределению, имеют одинаковый объем I, и т.п. Все бесконечное множество таких выборок образуют *генеральную совокупность* (называемую также *популяцией*).

4.4 Статистика

В математике слово *«статистика»* имеет два значения.

Во-первых, так называется раздел математики, в котором по выборке (результатам экспериментов) определяется вид распределения, из которого была извлечена эта выборка, оцениваются параметры этого распределения, проверяются гипотезы о виде этого распределения.

Второе значение слова «статистика» – это (измеримая) функция выборки. Поскольку элементы выборки суть случайные величины, то и статистика является случайной величиной. Назначение статистик – оценка параметров распределения, из которого извлечена выборка.

Примеры таких оценок приведены ниже.

4.5 Выборочное среднее и дисперсия

Выборочным средним называется статистика

$$\bar{x} = \frac{1}{I} \sum_{i=1}^{I} x_i$$

Для вычисления выборочной дисперсии используются две статистики:

— *смещенная* оценка:

$$s_m^2 = \frac{1}{I} \sum_{i=1}^{I} (x_i - \bar{x})^2$$

— *несмещенная* оценка:

$$s^2 = \frac{1}{I-1} \sum_{i=1}^{I} (x_i - \bar{x})^2$$

Смещенную оценку следует использовать если среднее значение известно заранее и не требует оценки. Аналогичным образом определяются выборочные моменты, например,

$$\overline{m_k} = \frac{1}{I-1}\sum_{i=1}^{I}(x_i - \bar{x})^k$$

является оценкой k-ого центрального момента.

Для вычисления выборочных статистик в Excel используют следующие стандартные функции: `AVERAGE` (`СРЗНАЧ`), `VAR` (`ДИСП`), `VARP` (`ДИСПР`), `STDEV` (`СТАНДОТКЛОН`), `STDEVP` (`СТАНДОТКЛОНП`).

`AVERAGE(x)`

Возвращает среднее значение выборки x, $\bar{x}$.

`VAR(x)`

Возвращает выборочную дисперсию выборки (несмещенную) x, s^2.

`VARP(x)`

Возвращает выборочную дисперсию выборки (смещенную) x, s_m^2.

`STDEV(x)`

Возвращает среднеквадратичное отклонение т.е. корень квадратный из выборочной дисперсии выборки x (несмещенной).

`STDEVP(x)`

Возвращает среднеквадратичное отклонение т.е. корень квадратный из выборочной дисперсии выборки x (смещенной).

4.6 Выборочные ковариации и корреляции

Если имеются две выборки $\mathbf{x} = (x_1, \ldots, x_I)$ и $\mathbf{y} = (y_1, \ldots, y_I)$, то можно рассчитать выборочные значения *ковариации* и *корреляции*. Ковариация c рассчитывается по формуле

$$c = \frac{1}{I} \sum_{i=1}^{I} (x_i - \bar{x})(y_i - \bar{y})$$

а коэффициент корреляции r по формуле

$$r = \sqrt{\frac{c}{s_x^2 s_y^2}}$$

В более общем случае, когда имеется матрица данных $\mathbf{X}$, размерностью I наблюдений на J переменных, то выборочная матрица ковариаций $\mathbf{C}_I$ между наблюдениями рассчитывается так:

$$\mathbf{C}_I = \mathbf{X}\mathbf{X}^{\mathrm{t}}$$

Выборочная матрица ковариаций $\mathbf{C}_J$ между переменными так:

$$\mathbf{C}_J = \mathbf{X}^{\mathrm{t}}\mathbf{X}$$

Для вычисления парных ковариаций в Excel используют следующие стандартные функции: `COVAR` (`КОВАР`), `CORREL` (`КОРРЕЛ`).

```
COVAR(x, y)
```

Возвращает выборочную ковариацию между выборками x и y.

```
CORREL(x, y)
```

Возвращает выборочный коэффициент корреляции между выборками x и y.

4.7 Вариационный ряд и порядковые статистики

Исходную выборку $(x_1, \dots, x_I)$ можно упорядочить в порядке неубывания:

$$x(1) \le x(2) \le \cdots \le x(i) \le \cdots \le x(I)$$

и получить т.н. *вариационный* ряд.

Элементы этого ряда являются *порядковыми* статистиками. Центральный элемент ряда (а если I – четное, то полусумма двух центральных) является выборочной оценкой медианы

$$\text{median}(\mathbf{x}) = \begin{cases} x(k+1), I = 2k+1 \\ 0.5(x(k) + x(k+1)), I = 2k \end{cases}$$

Аналогичным способом строятся оценки квартилей и процентилей.

Размахом выборки называется величина

$$x(I) - x(1)$$

Интерквартильным размахом выборки $\mathbf{x}$ называется величина

$$IQR(\mathbf{x}) = \hat{x}(0.75) - \hat{x}(0.25)$$

являющаяся разностью выборочных квартилей для $P = 0.75$ и $P = 0.25$.

Для вычисления порядковых статистик в Excel используют следующие стандартные функции: `MEDIAN` (`МЕДИАНА`), `QUARTILE` (`КВАРТИЛЬ`), `PERCENTILE` (`ПЕРСЕНТИЛЬ`).

```
MEDIAN(x)
```

Возвращает выборочную медиану для выборки x.

```
QUARTILE(x, quart=0|1|2|3|4)
```

Возвращает выборочный квартиль для выборки x в зависимости от значения аргумента `quart`: 0 — минимальное значение 1 — первый квартиль (25-ый перцентиль) 2 — значение медианы (50-ый перцентиль) 3 — третий квартиль (75-ый перцентиль) 4 — максимальное значение

```
PERCENTILE(x, k)
```

Возвращает k-ый выборочный перцентиль для выборки x. Значения аргумента: $0 \leq k \leq 1$.

4.8 Выборочная функция распределения

Выборочная (или *эмпирическая*) функция распределения – это неубывающая функция $F_I(x)$, которая равна нулю при $x < x(1)$ и равна 1 при $x \geq x(I)$. Между этими двумя точками функция $F_I(x)$ ступенчато возрастает на величину $1/I$ каждый раз при переходе через следующую точку $x(i)$:

$$F_I(x) = \frac{x(i) \leq x}{I}$$

Выборочная функция распределения имеет важное теоретическое значение, т.к. при увеличении объема выборки I эмпирическая функция сходится к истинной функции распределения. Однако в практических приложениях чаще используется гистограмма.

Для построения гистограммы область изменения выборочных значений $[x(1), x(I)]$ разбивается на R частей равного размера. Затем подсчитывается, сколько элементов выборки попало в каждую из этих областей: $I_1 + I_2 + \cdots + I_R = I$. После этого частоты $F_r = I_r/I$ откладывают на ступенчатом графике, аналогичном показанному на Рис. 4.1.

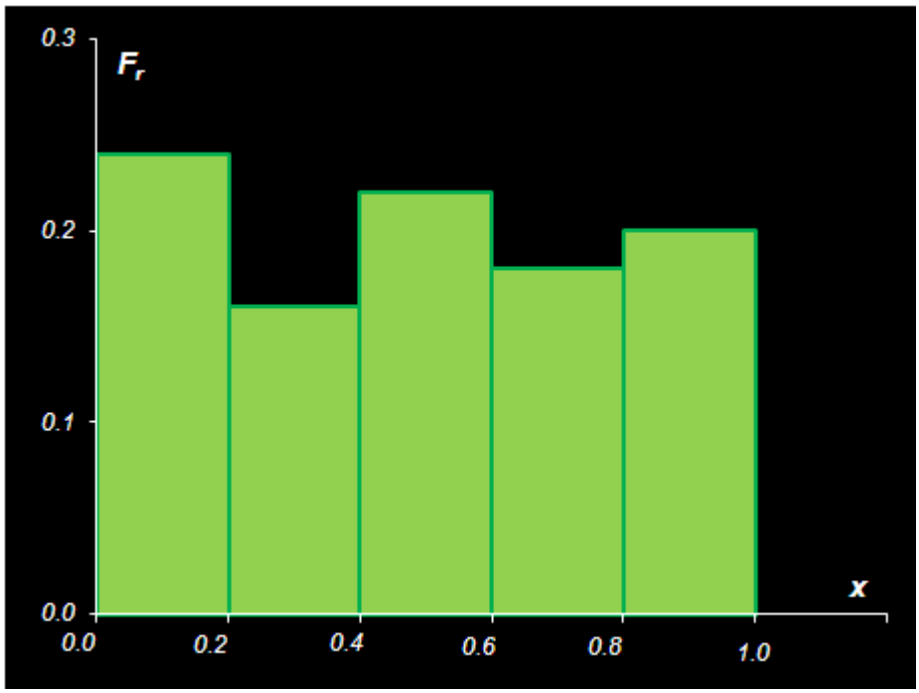


Рис. 4.1. Гистограмма

Для построения гистограмм в Excel применяется стандартная функция `FREQUENCY` (`ЧАСТОТА`).

```
FREQUENCY (data_array, bins_array)
```

Возвращает число попаданий значений `data_array` в интервалы, заданные аргументом `bins_array`. Эта функция возвращает вертикальный массив, и она должна вводится как формула массива – с помощью комбинации клавиш `CTRL+SHIFT+ENTER`. Количество элементов в возвращаемом массиве на единицу больше числа элементов в массиве `bins_array`. Дополнительный элемент содержит количество значений из `data_array` больших, чем максимальное значение в массиве `bins_array`.

	A	B	C	D	E	F	G	H
2		X			Bins	l_r	F_r	P_r
3	1	0.823		1	0.0	=FREQUENCY(X,Bins)		
4	2	0.583		2	0.2	12	0.24	0.2
5	3	0.886		3	0.4	8	0.16	0.2
6	4	0.934		4	0.6	11	0.22	0.2
7	5	0.847		5	0.8	9	0.18	0.2
8	6	0.179		6	1.0	10	0.2	0.2
9	7	0.329				0	0	
10	8	0.485						

Рис. 4.2. Пример использования функции FREQUENCY

4.9 Метод моментов

Все рассмотренные выше методы анализа выборок не учитывали конкретный вид распределения, из которого эта выборка была извлечена. Такие способы оценивания называются *непараметрическими.*

Теперь мы рассмотрим типичный *параметрический* метод моментов. Пусть выборка $\mathbf{x} = (x_1, \dots, x_I)$ подчиняется функции распределения

$$x_i \sim F(x|\mathbf{p})$$

которая известна с точностью до значений параметров $\mathbf{p} = (p_1, \dots, p_M)$. Для оценки параметров вычислим M выборочных моментов и приравняем их соответствующим теоретическим значениям. Получится система нелинейных уравнений

$$\begin{cases} m_1(p_1, \dots, p_M) = \overline{m}_1 \\ m_2(p_1, \dots, p_M) = \overline{m}_2 \\ \dots \\ m_M(p_1, \dots, p_M) = \overline{m}_M \end{cases}$$

из которой нужно численно найти значения оценок параметров $\mathbf{p}$.

Для примера рассмотрим случайную величину $X = aY$, где величина Y распределена по хи-квадрат

$$Y \sim \chi^2(N)$$

По выборке $\mathbf{x} = (x_1, \dots, x_I)$ нужно найти оценки двух неизвестных параметров a и N.

Из раздела 2.6 следует:

$$\mathrm{E}(X) = a\mathrm{E}(Y) = aN$$

$$\mathrm{V}(X) = a^2\mathrm{V}(Y) = 2a^2N$$

Поэтому:

$$\hat{a}_{MM} = \frac{\mathrm{V}(X)}{2\mathrm{E}(X)} = \frac{s^2}{2\overline{m}}$$

$$\hat{N}_{MM} = \frac{\mathrm{E}(X)}{a} = \frac{2\overline{m}^2}{s^2}$$

4.10 Метод максимума правдоподобия

Самый популярный способ параметрического оценивания – это метод максимума правдоподобия. Учитывая, что каждый элемент выборки $\mathbf{x} = (x_1, \dots, x_I)$ имеет одну и ту же плотность вероятности $f(x_i|\mathbf{p})$, совместная плотность всей выборки имеет вид:

$$L(\mathbf{x}|\mathbf{p}) = f(x_1|\mathbf{p}) \times f(x_2|\mathbf{p}) \times \cdots \times f(x_I|\mathbf{p}) = \prod_{i=1}^{I} f(x_i|\mathbf{p})$$

Функция $L(\mathbf{x}|\mathbf{p})$ называется *функцией правдоподобия выборки*. Она зависит от двух групп переменных – выборочных значений $\mathbf{x} = (x_1, \ldots, x_I)$, известных из эксперимента, и параметров $\mathbf{p} = (p_1, \ldots, p_M)$, которые предстоит оценить.

В качестве оценок берутся такие значения параметров $\mathbf{p}$, при которых функция правдоподобия (или ее логарифм) имеет максимум

$$\hat{\mathbf{p}} = \arg\max_{\mathbf{p}} \big(\ln(L(\mathbf{x}|\mathbf{p})) \big)$$

Рассмотрим, для примера, оценки параметров нормального распределения $N(\mu, \sigma^2)$. Вспоминая, что для независимых случайных величин многомерную функцию плотности распределения можно представить в виде произведения одномерных функций и зная выражение для функции плотности нормального распределения можно вывести следующее:

$$L(\mathbf{x}|\mu, \sigma^2) = (\sigma\sqrt{2\pi})^{-I} \exp\Big(- \frac{\sum_{i=1}^{I}(x_i - \mu)^2}{\sigma^2} \Big)$$

Максимум этой функции достигается при следующих значениях параметров:

$$\hat{\mu}_{ML} = \frac{1}{I} \sum_{i=1}^{I} x_i$$

$$\hat{\sigma}^2_{ML} = \frac{1}{I} \sum_{i=1}^{I} (x_i - \hat{\mu}_{ML})^2$$

Таким образом, для нормального распределения оценки МП совпадают с выборочными оценками приведенными нами ранее.

5 Свойства оценок

5.1 Состоятельность

Любая оценка $p(\mathbf{x})$ параметра p есть статистика, т.е. случайная величина. И как всякая случайная величина она обладает собственной функцией распределения, математическим ожиданием, дисперсией и т.д. Все эти характеристики позволяют сравнивать разные оценки, судить об их свойствах и качествах. Ниже следует краткий обзор основных свойств оценок.

Оценка $p(\mathbf{x})$ называется состоятельной, если она сходится по вероятности к значению оцениваемого параметра p при безграничном возрастании объема выборки I. Точнее, статистика $p(\mathbf{x})$ является состоятельной оценкой параметра p тогда и только тогда, когда для любого положительного числа ϵ справедливо

$$\lim_{I\to\infty} \Pr\big(|p(\mathbf{x}) - p| > \epsilon\big) = 0$$

Большинство оценок, используемых в практических приложениях, являются состоятельными. ## Смещенность

Оценка $p(\mathbf{x})$ называется несмещенной, если:

$$\mathrm{E}[p(\mathbf{x})] = p$$

Смещенные оценки часто встречаются в приложениях. Например, МП-оценка дисперсии нормального распределения является смещенной:

$$\mathrm{E}(\sigma^2_{ML}(\mathbf{x})) = (1 - 1/I)\sigma^2$$

Для несмещенных оценок мерилом их точности является дисперсия $\mathrm{V}[p(\mathbf{x})]$ – чем она меньше, тем лучше. Для смещенных оценок нужно использовать математическое ожидание квадрата смещения.

$$d(\mathbf{x}) = \mathrm{E}[(p(\mathbf{x}) - p)^2]$$

Имеет место формула:

$$d(\mathbf{x}) = \mathrm{V}[p(\mathbf{x})] + \mathrm{E}[(p(\mathbf{x})] - p)^2$$

5.2 Эффективность

Несмещенная оценка называется эффективной, если она имеет наименьшую возможную дисперсию. Оценки нормального распределения являются эффективными, но вот выборочная оценка медианы (см. раздел 4.7) таковой не является – она менее эффективно оценивает μ, чем выборочное среднее.

Смещенные оценки могут оказаться более точными, чем несмещенные. Это означает, что часто можно построить такие смещенные оценки, для которых квадрат ошибки меньше, чем наименьшая эффективная дисперсия. На этом принципе основаны такие методы оценивания как PCR, PLS и др.## Робастность

Робастность оценки – это важная характеристика, которая, однако, плохо поддается формализации.

Оценка $p(\mathbf{x})$ называется *робастной*, если она устойчива к наличию выбросов в выборке.

Как правило, эффективные оценки являются менее робастными, чем неэффективные. Выбирая более устойчивую оценку, мы расплачиваемся за это эффективностью.

Для нормального распределения робастной оценкой среднего значения является медиана, а для СКО можно использовать MAD-оценку

$$s_{MAD} = 1.4826\mathrm{median}\big(|\mathbf{x} - \mathrm{median}(\mathbf{x})\big)$$

На листе `Robust` приведено сравнение классических и робастных оценок для выборки из стандартного нормального распределения $N(0,1)$, в которой первый элемент заменен на выброс – случайную величину из распределения $N(0,100)$.

5.3 Нормальная выборка

Если выборка $\mathbf{x} = (x_1, \dots, x_I)$ извлечена из нормального распределения

$$x_i \sim N(\mu, \sigma^2)$$

и оценки $\bar{x}$, s^2, определены формулами, то выполняются следующие утверждения.

$$\sqrt{I}\frac{\bar{x} - \mu}{\sigma} \sim N(0,1)$$

т.е. имеет стандартное нормальное распределение;

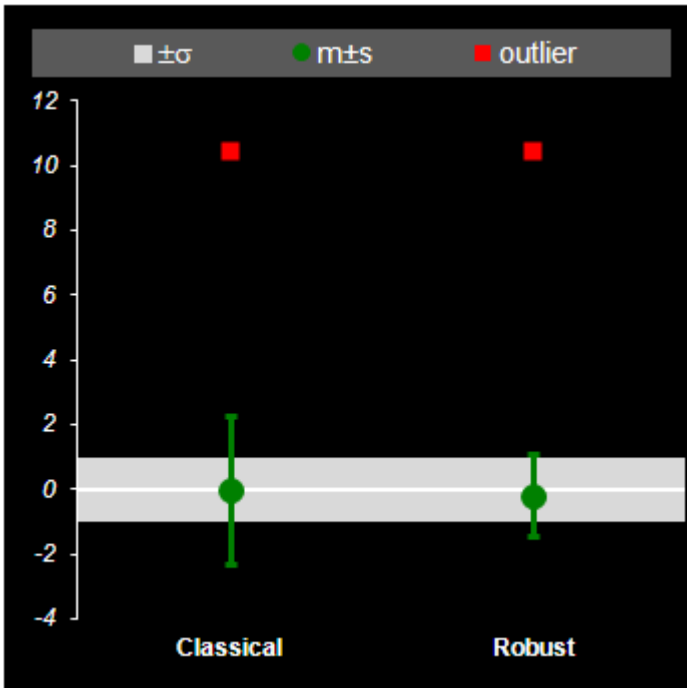


Рис. 5.1. Обычные и робастные оценки

$$I\frac{s_m^2}{\sigma^2} \sim \chi^2(I)$$

т.е.имеет распределение хи-квадрат с I степенями свободы;

$$I\frac{s^2}{\sigma^2} \sim \chi^2(I-1)$$

т.е.имеет распределение хи-квадрат с $I-1$ степенями свободы;

$$\sqrt{I-1}\frac{\bar{x}-\mu}{\sigma} \sim T(I-1)$$

т.е. имеет распределение Стьюдента с $I-1$ степенью свободы.

6 Доверительное оценивание

6.1 Доверительная область

Во многих случаях, помимо точечных оценок неизвестных параметров распределения, желательно указать область, в которой истинные значения этих параметров содержатся с заданной вероятностью. Такая область называется *доверительной*.

Дадим точное определение. Пусть выборка $\mathbf{x} = (x_1, \dots, x_I)$ подчиняется функция распределения $F(x|\mathbf{p})$, т.е.

$$x_i \sim F(x|\mathbf{p})$$

которая известна с точностью до значений параметров $\mathbf{p} = (p_1, \dots, p_M)$. Статистика $P(\mathbf{x}) \in R^M$ называется доверительной областью, соответствующей доверительной вероятности γ, если:

$$\Pr \mathbf{p} \in P(\mathbf{x}) \geq \gamma$$

6.2 Доверительный интервал

Часто для каждого параметра p_m строится своя одномерная область – *доверительный интервал*.

Границы доверительного интервала – это две статистики $p^-(\mathbf{x})$ и $p^+(\mathbf{x})$, такие, что

$$\Pr p^-(\mathbf{x}) \leq p \leq p^+(\mathbf{x}) \geq \gamma$$

Для односторонних доверительных интервалов соответствующая граница заменяется на $-\infty$, 0, или $+\infty$.

В большинстве практических случаев доверительные интервалы строятся для (асимптотически) нормальных выборок с помощью соотношений, приведенных в разделе 5.3.

6.3 Пример построения интервала

Приведем пример построения доверительного интервала.

Пусть имеется выборка $\mathbf{x} = (x_1, \dots, x_I)$ из нормального распределения $N(\mu, \sigma^2)$ с известной дисперсией σ^2. Построим доверительный интервал для параметра μ – математического ожидания.

Из раздела 5.3 следует, что

$$\Pr\left\{\Phi^{-1}(1-\alpha_1) \le \frac{\bar{x}-\mu}{\sigma}\sqrt{I} \le \Phi^{-1}(1-\alpha_2)\right\} = \alpha_1 + \alpha_2 - 1$$

где Φ^{-1} – квантиль стандартного нормального распределения, поэтому

$$\Pr\left\{\bar{x} - \frac{\sigma}{\sqrt{I}}\Phi^{-1}(\alpha_2) \le \mu \le \bar{x} + \frac{\sigma}{\sqrt{I}}\Phi^{-1}(\alpha_1)\right\} = \alpha_1 + \alpha_2 - 1$$

Для построения симметричного доверительного интервала с доверительной вероятностью γ, положим $\alpha_1 = \alpha_2 = 0.5(1+\gamma)$. Для построения односторонних доверительных интервалов, положим $\alpha_1 = 1$, $\alpha_2 = \gamma$, или $\alpha_1 = \gamma, \alpha_2 = 1$. ## Интервалы для нормального распределения

Используя соотношения, приведенные в разделе 5.3, можно построить доверительный интервал для параметров нормального распределения $N(\mu, \sigma^2)$.

Пусть имеются оценки $\bar{x}$, s^2, s_m^2, тогда выполняются следующие утверждения.

Доверительный интервал для среднего значения μ при неизвестной дисперсии $$⊠^21 имеет вид:

$$\Pr\left\{\bar{x} - \frac{s}{\sqrt{I}}T^{-1}(\alpha_2|I-1) \le \mu \le \bar{x} + \frac{s}{\sqrt{I}}T^{-1}(\alpha_1|I-1)\right\} = \alpha_1 + \alpha_2 - 1$$

где $T^{-1}(\alpha|I-1)$ – квантиль распределения Стьюдента с $I-1$ степенями свободы.

Доверительный интервал для дисперсии σ^2 при известном среднем значении μ имеет вид:

$$\Pr\left\{I\frac{s_m^2}{\chi^{-2}(\alpha_2|I)} \le \sigma^2 \le I\frac{s_m^2}{\chi^{-2}(\alpha_1|I)} \le \right\} = \alpha_1 + \alpha_2 - 1$$

где $\chi^{-2}(\alpha|I)$ – квантиль распределения хи-квадрат с I степенями свободы.

Доверительный интервал для дисперсии σ^2 при неизвестном среднем значении μ имеет вид:

$$\Pr\left\{I\frac{s^2}{\chi^{-2}(\alpha_2|I-1)} \le \sigma^2 \le I\frac{s^2}{\chi^{-2}(\alpha_1|I-1)} \le \right\} = \alpha_1 + \alpha_2 - 1$$

где $\chi^{-2}(\alpha|I-1)$ – квантиль распределения хи-квадрат с $I-1$ степенями свободы.

7 Проверка гипотез

7.1 Постановка задачи

Статистической гипотезой называется непротиворечивое утверждение, касающееся вида распределения имеющейся выборки.

Основная гипотеза, нуждающаяся в проверке называется *нулевой* или *нуль-гипотезой*. Любая другая гипотеза, относительно которой проверяют нуль-гипотезу, называется *альтернативой*. Например: пусть имеется выборка из распределения хи-квадрат с N степенями свободы. Нуль-гипотеза состоит в том, что –

$$H_0 : N = 2$$

альтернатива –

$$H_1 : N > 2$$

На практике альтернативу часто опускают, формулируя только нуль-гипотезу.

Гипотеза называется *простой*, если она однозначно определяет функцию распределения выборки. В противном случае гипотеза называется сложной. В примере: H_0 – это простая гипотеза, а H_1 – это сложная альтернатива.

Гипотезы бывают *параметрическими*, когда вид распределения известен заранее, с точностью до численных значений его параметров – как в примере выше. Кроме того, гипотезы могут быть непараметрическими.

Например: пусть имеется выборка из неизвестного распределения F. Нуль-гипотеза состоит в том, что – $H_0 : F$ – это равномерное распределение.

7.2 Проверка гипотез

Метод проверки статистической гипотезы называется статистическим критерием. Он строится на основе имеющейся выборки $\mathbf{x} = (x_1, \dots, x_I)$ с помощью измеримой функции $S(\mathbf{x})$, называемой статистикой критерия. В пространстве значений статистики $S(\mathbf{x})$ выбирается область C, называемая критической. Если $S(\mathbf{x}) \in$, то гипотезу отклоняют (отвергают), в противном случае – принимают.

Статистика $S(\mathbf{x})$ должна быть устроена особым образом – так, чтобы ее распределение не зависело от неизвестных параметров распределения выборки $\mathbf{x}$. Кроме того функция распределения $S(\mathbf{x})$ должна быть табулирована заранее.

В большинстве практических приложений статистика $S(\mathbf{x})$ строится из соображений нормальности.## Ошибки 1-го и 2-го родов

Проверка статистической гипотезы не дает ее логического подтверждения или опровержения. Проверка только утверждает, что “имеющиеся данные (не) противоречат» выдвинутому предположению”. Поэтому при проверке статистической гипотезы возможны случайные ошибки, которые могут быть двух родов.

Ошибка 1-го рода происходит тогда, когда нуль-гипотеза верна, но отвергается согласно критерию.

Ошибка 2-го рода происходит тогда, когда нуль-гипотеза не верна, но принимается согласно критерию.

Вероятность ошибки первого рода называется [уровнем значимости]http://ru.wikipedia.org/wiki/Уровень_-значимости и обозначается α.

Обычно уровень значимости выбирается равным 0.01, 0.05, или 0.1 и по этому значению подбирают критическую область C_α.## Пример проверки гипотезы

Пусть имеется выборка $\mathbf{x} = (x_1, \dots, x_I)$ из нормального распределения –

$$x_i \sim N(\mu, \sigma^2)$$

с *известной дисперсией* σ^2 и неизвестным средним μ.

Проверяется простая нуль-гипотеза:

$$H_0 \ : \ \mu = 0$$

Альтернативу мы сформулируем позже.

В качестве статистического критерия возьмем функцию

$$S(\mathbf{x}) = \sqrt{I}\frac{\bar{x}}{\sigma}$$

которая при $\mu = 0$ подчиняется стандартному нормальному распределению –

$$S \sim N(0, 1)$$

При заданном уровне значимости α критическая область определяется условием –

$$\Pr\{|S| > C_\alpha\} = \alpha$$

Поэтому

$$C_\alpha = \Phi^{-1}(1-\alpha/2)$$

Введем теперь альтернативную гипотезу –

$$H_1 \ : \ \mu = a$$

и найдем величину ошибки 2-го рода. Ее величина

$$\beta = \Pr\{|S| < C_\alpha | \mu = a\}$$

рассчитывается при условии

$$S \sim N(, 1)$$

Поэтому,

$$\beta = \Phi(C_\alpha - a) - \Phi(-C_\alpha - a)$$

На листе `Hypothesis` приведены расчеты этого примера.

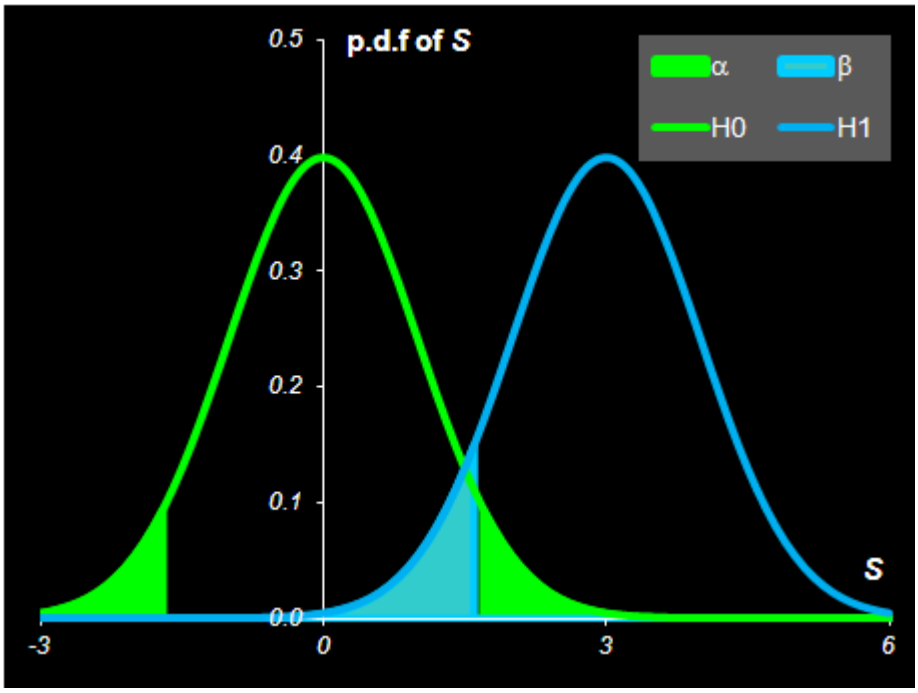


Рис. 7.1. Ошибки 1-го и 2-го родов при проверке гипотез

7.3 Критерий согласия хи-квадрат

Критерий согласия хи-квадрат проверяет соответствие между теоретическими вероятностями $P_1, P_2, \ldots$ и их эмпирическими частотными оценками $I_1/I, I_2/I, \ldots$.

Для примера рассмотрим выборку $\mathbf{x} = (x_1, \ldots, x_I)$ из неизвестного распределения –

$$x_i \sim F(x)$$

Нуль гипотеза состоит в конкретизации этого распределения, т.е. в утверждении типа «F – это нормальное распределение с нулевым средним и дисперсией равной 2»

В соответствие с выбранным гипотетическим распределением, область изменения случайной величины X, разбивается на R классов (корзин) и рассчитываются теоретические вероятности $P_1, P_2, \ldots, P_R$ попадания в каждую из корзин. С другой стороны определяется, сколько элементов выборки попало в каждую из этих корзин – $I_1, I_2, \ldots, I_R$ и вычисляются эмпирические вероятности $F_r = I_r/I$.

Статистикой критерия согласия служит случайная величина

$$S = \sum_{r=1}^{R} \frac{(I_r - IP_r)^2}{IP_r} = I \sum_{r=1}^{R} \frac{(F_r - P_r)^2}{P_r}$$

которая при $I \to \infty$ стремится к распределению хи-квадрат с $R-1$ степенями свободы. Число и размеры корзин надо выбирать так, чтобы

$$IP_r > 6$$

Критическая область на уровне значимости α определяется условием –

$$S > \chi^{-2}(1-\alpha|R-1)$$

Критерий согласия хи-квадрат можно применять и в том случае, когда теоретическое распределение $F(x|\mathbf{p})$ известно с точностью до неизвестных параметров $\mathbf{p} = (p_1, \ldots, p_M)$. Эти параметры предварительно оцениваются по той же выборке $\mathbf{x}$ и подставляются в функцию $F(x|\mathbf{p})$. В этом случае следует изменить число степеней свободы на $R-M-1$.

Для проверки согласия по критерию хи-квадрат в Excel применяется стандартная функция `CHITEST` (`ХИ2ТЕСТ`):

```
CHITEST(actual_range, expected_range)
```

Вычисляет статистику S приведеную выше используя `actual_range=(I1, I2,...,IR)` и `expected_range=(IP1, IP2,...,IPR)`. Возвращает вероятность $P = 1-\chi^2(S|R-1)$.

Для принятия гипотезы на уровне значимости α необходимо, чтобы $P > 1-\alpha$.

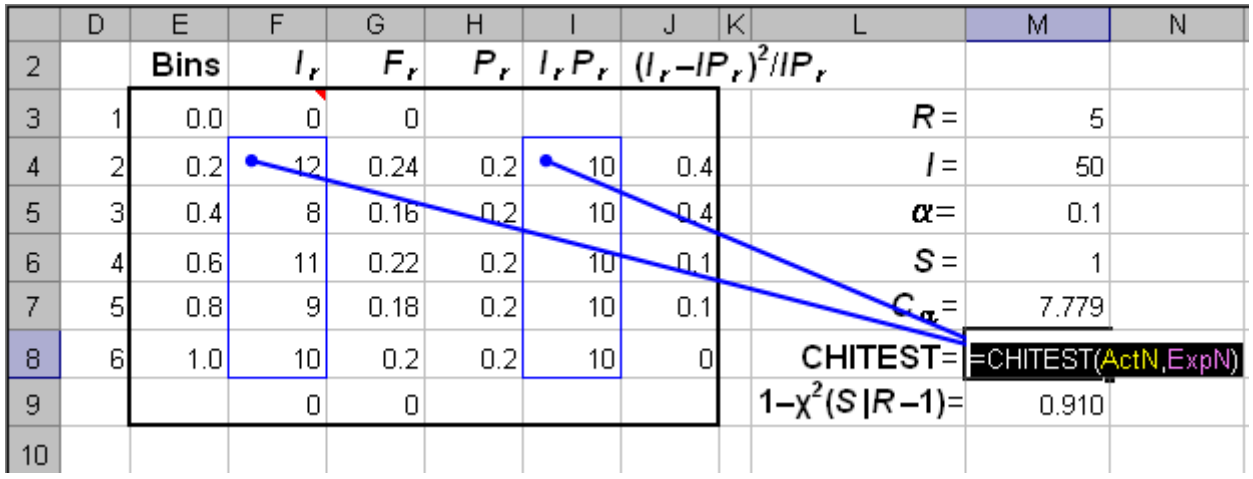


Рис. 7.2. Пример проверки критерия согласия хи-квадрат

7.4 F-критерий

Этот критерий применяется для проверки нуль-гипотезы о равенстве дисперсий в двух нормальных выборках: $\mathbf{x} = (x_1, \dots, x_I)$ и $\mathbf{y} = (y_1, \dots, y_J)$. Пусть s_x^2, s_y^2 – суть оценки выборочных дисперсий.

Если $s_x^2 > s_y^2$, то обозначим:

$$s_1^2 = s_x^2, N_1 = I - 1$$

$$s_2^2 = s_y^2, N_2 = J - 1$$

Иначе:

$$s_1^2 = s_y^2, N_1 = J - 1$$

$$s_2^2 = s_x^2, N_2 = I - 1$$

Статистикой F-критерия служит случайная величина

$$S = \frac{s_1^2}{s_2^2} \sim F(N_1, N_2)$$

которая подчиняется распределению Фишера (раздел 3.6) с N_1, N_2 степенями свободы.

Критическая область на уровне значимости α определяется условием:

$$S > F^{-1}(1-\alpha|N_1, N_2)$$

F-критерий очень чувствителен к нарушению предположения о нормальности распределений выборок, поэтому его не рекомендуется применять в практических приложениях.

Для проверки F-критерия в Excel применяется стандартная функция `FTEST` (`ФТЕСТ`):

```
FTEST(x, y)
```

Возвращает вероятность $P = 2[1-F(S|N_1, N_2)]$. Для принятия гипотезы на уровне значимости α необходимо, чтобы $P > 2\alpha$.

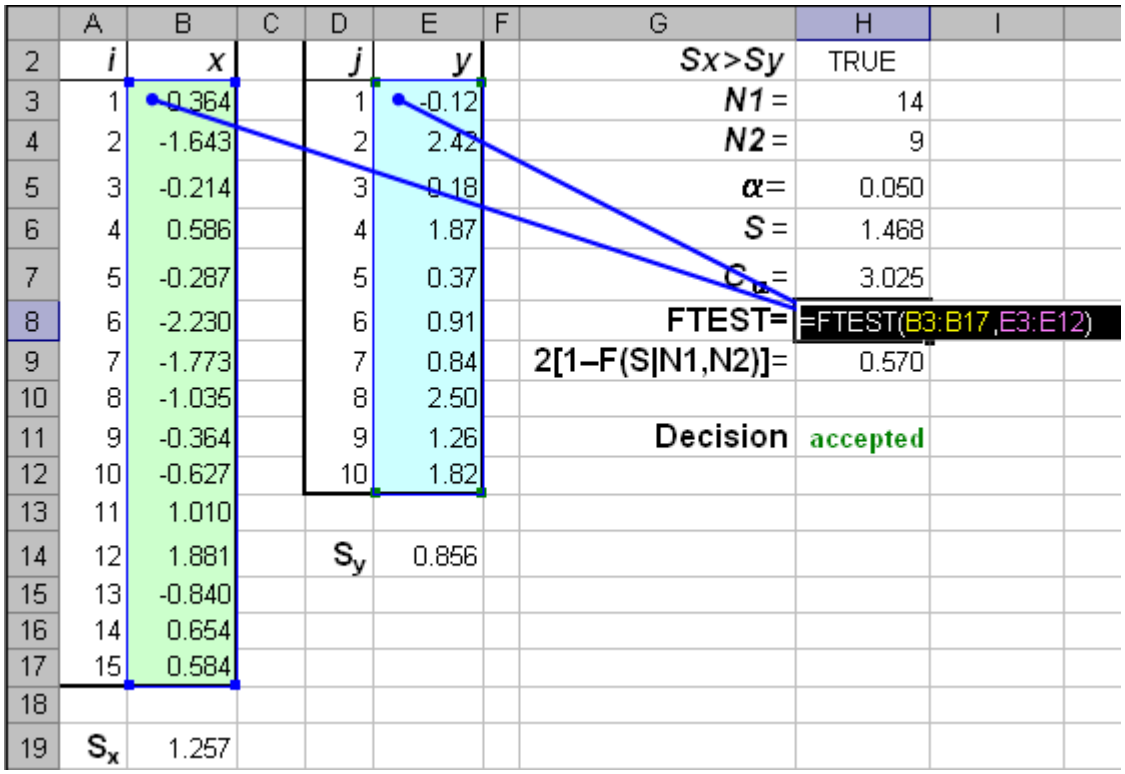


Рис. 7.3. Пример проверки F-критерия

8 Регрессия

8.1 Простейшая регрессия

В простейшей постановке в регрессионном анализа рассматриваются две выборки детерминированных величин $\mathbf{x} = (x_1, \dots, x_I)$ и случайных величин $\mathbf{y} = (y_1, \dots, y_I)$.

Набор $\mathbf{x}$ называется *предикторами*, а набор $\mathbf{y}$ – *откликами*. Предполагается, что между этими величинами существует линейная связь вида

$$y_i = ax_i + b + \varepsilon_i$$

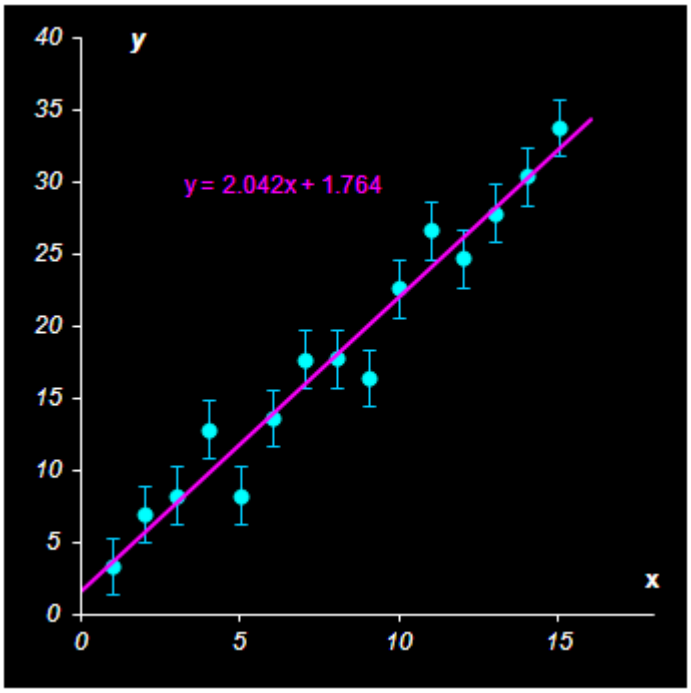
где a и b – неизвестные параметры, а ε_i – ошибки, т.е. некоррелированные случайные величины, имеющие нулевое мат.ожидание и неизвестную дисперсию σ^2.

Если дополнительно предположить, что ошибки распределены нормально

$$\varepsilon_i \sim N(0, \sigma^2)$$

то оценивание параметров a и b методом максимального правдоподобия сведется к поиску минимума следующей суммы квадратов

$$Q(a, b) = \sum_{i=1}^{I} (y_i - ax_i - b)^2$$



{#fig20}

8.2 Метод наименьших квадратов

На практике допущение о нормальности является избыточным и метод наименьших квадратов применяют и в случае ошибок произвольного вида.

Минимум суммы Q достигается в точке:

$$\hat{b} = \frac{\sum_{i=1}^{I}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{I}(x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Дисперсии оценок параметров равны:

$$\mathrm{V}(\hat{a}) = \frac{\sigma^2}{I}\frac{\sum_{i=1}^{I} x_i^2}{\sum_{i=1}^{I}(x_i - \bar{x})^2}$$

$$\mathrm{V}(\hat{b}) = \frac{\sigma^2}{I\sum_{i=1}^{I}(x_i - \bar{x})^2}$$

Оценка параметра σ^2 равна:

$$s^2 = \frac{Q(\hat{a}, \hat{b})}{I-2} = \frac{1}{I-2}\sum_{i=1}^{I}(y_i - \hat{a}x_i - \hat{b})^2$$

Если допущение о нормальности верно, то выполняются следующие соотношения:

$$(I-2)\frac{s^2}{\sigma^2} \sim \chi^2_{I-2}$$

$$\hat{a} \sim N\big(a, \mathrm{V}(\hat{a})\big)$$

$$\hat{b} \sim N\big(b, \mathrm{V}(\hat{b})\big)$$

Кроме того, оценки параметров a и b независимы от оценки σ^2, что позволяет строить для оценок параметров доверительные интервалы с помощью статистики Стьюдента (раздел 3.5).

Для вычисления параметров регрессии в Excel используют две стандартные функции: `SLOPE` (`НАКЛОН`) и `INTERCEPT` (`ОТРЕЗОК`):

```
SLOPE (known_y=y, known_x=x)
```

Возвращает оценку параметра a.

```
INTERCEPT (known_y=y, known_x=x)
```

Возвращает оценку параметра b.

	A	B	C	D	E	F	G	H
2	a=	2.000			a^{hat}=	1.920		
3	b=	1.000			b^{hat}=	=INTERCEPT(B7:B21,A7:A21)		
4	σ=	2			b^{hat}=	2.089		
5								
6	x	y	y^{hat}	$y-y^{hat}$				
7	1	1.752	2.940	-1.189				
8	2	4.412	4.860	-0.448				
9	3	4.863	6.780	-1.916				
10	4	11.728	8.699	3.028				
11	5	8.459	10.619	-2.160				
12	6	14.870	12.539	2.332				
13	7	13.119	14.458	-1.339				
14	8	18.419	16.378	2.041				
15	9	21.161	18.298	2.863				
16	10	20.371	20.217	0.154				
17	11	22.834	22.137	0.697				
18	12	19.915	24.057	-4.142				
19	13	26.338	25.976	0.361				
20	14	28.054	27.896	0.158				
21	15	29.375	29.816	-0.440				
22								

{#fig-21}

8.3 Множественная регрессия

Естественным обобщением простой однофакторной регрессии является множественная регрессия –

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \boldsymbol{\varepsilon}$$

в которой рассматривается связь между вектором откликов $\mathbf{y}$ и матрицей предикторов $\mathbf{X}$. Обычное предположение относительно ошибок состоит в том, что:

$$E(\varepsilon_i) = 0, \mathrm{cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

где $\mathbf{I}$ – единичная $(I \times I)$ матрица.

Если в задаче имеется I наблюдений и J переменных, то матрица $\mathbf{X}$ имеет размерность $I \times J$. Цель регрессионного анализа – найти оценки неизвестных коэффициентов $\mathbf{a} = (a_1, \ldots, a_J)^{\mathrm{t}}$, такие, которые минимизируют сумму квадратов остатков:

$$Q(\mathbf{a}) = \|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2 = \sum i = 1^I |y_i - \sum_{j=1}^{J} x_{ij} a_j|^2$$

В обычном методе наименьших квадратов (МНК) предполагается, что матрица $\mathbf{X}^t\mathbf{X}$ обратима. Тогда минимум $Q(\mathbf{a})$ достигается при:

$$\hat{a} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

Оценка параметра σ^2 равна:

$$s^2 = Q(\hat{\mathbf{a}})/(I-J)$$

Матрица ковариаций оценок равна

$$\mathbf{C} = s^2(\mathbf{X}^t\mathbf{X})^{-1}$$

Для построения множественной регрессии в Excel используют две стандартные функции: `TREND` (`ТЕНДЕНЦИЯ`) и `LINEST` (`ЛИНЕЙН`). Регрессия строится в следующем виде:

$$y = b + m_1 x_1 + \cdots + m_J x_j$$

```
TREND(known_y [,known_x] [,new_x] [,const])
```

`known_y` – вектор известных значений откликов $\mathbf{y}$ (калибровочный набор);

`known_x` – необязательный аргумент. Матрица известных значений предикторов $\mathbf{X}$ (калибровочный набор);

`new_x` – необязательный аргумент. Матрица новых значений предикторов $\mathbf{X}_{new}$ для которых вычисляются и выводятся значения откликов (проверочный набор);

`const` – необязательный аргумент. Логическое значение, которое указывает, требуется ли, чтобы параметр b был равен нулю. Если `const` имеет значение `TRUE` или опущено, то b вычисляется обычным образом, иначе $b = 0$.

Примечания:

- Вектор `known_y` должен занимать один столбец, тогда каждый столбец матрицы массива `known_x` интерпретируется как отдельная переменная;
- Если аргумент `known_x` опущен, то предполагается, что это вектор чисел 1; 2; 3; … такого же размера, как и `known_y`;
- Матрица новых значений `new_x` должна иметь столько же столбцов (переменных), как и матрица `known_x`;
- Если аргумент `new_x` опущен, то предполагается, что он совпадает с массивом `known_x`. Результат является вектором, в котором число строк равно числу строк в массиве `new_x`.

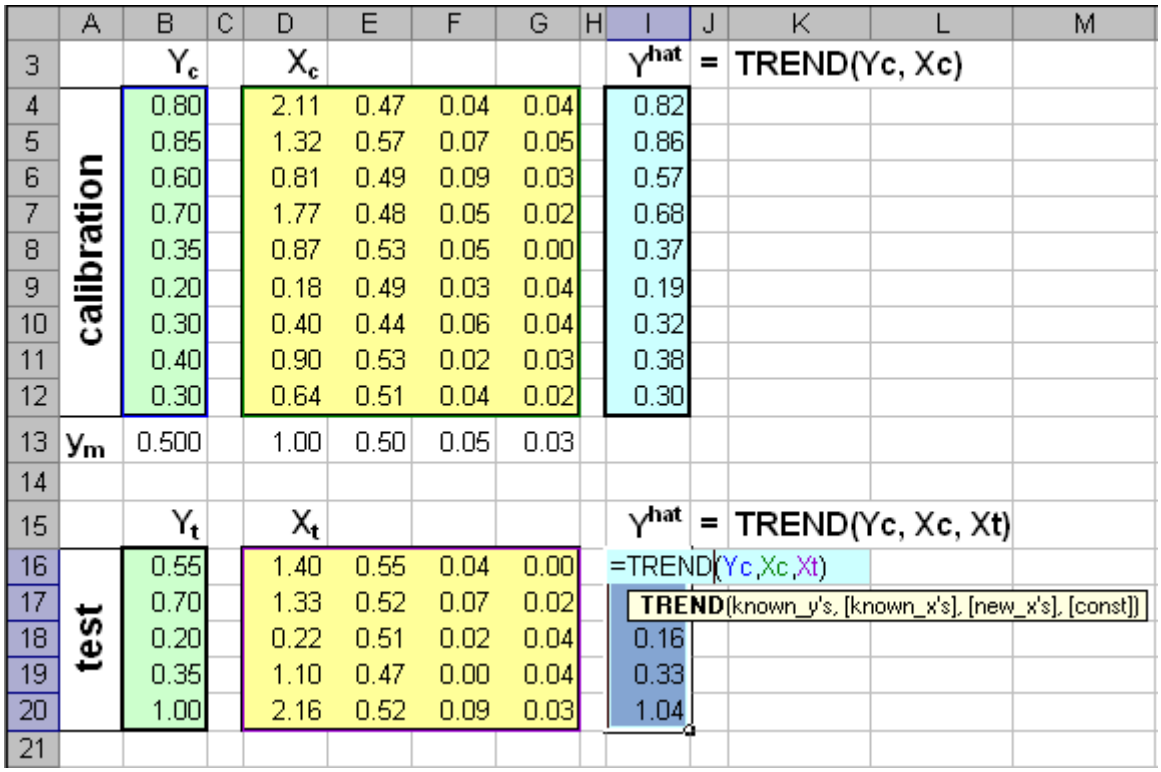


Рис. 8.1. Пример использования функции TREND

Функция `TREND` является функцией массива и ее ввод должен завершаться нажатием комбинации `CTRL+SHIFT+ENTER`.

Функция `LINEST` (`ЛИНЕЙН`) дополняет функцию `TREND` и выводит некоторые статистические значения, связанные с регрессией:

```
LINEST(known_y [,known_x] [,new_x] [,const] [,stats])
```

Первые четыре аргументы аналогичны параметрам функции `TREND`. Необязательный логический аргумент `stats` указывает, нужно ли выводить дополнительные статистические значения. Если `stats` имеет

значение FALSE или опущено, то выводятся только оценки коэффициентов $b, m_1, \ldots, m_J$. Иначе выводится целая таблица как показано на рисунке 8.2.

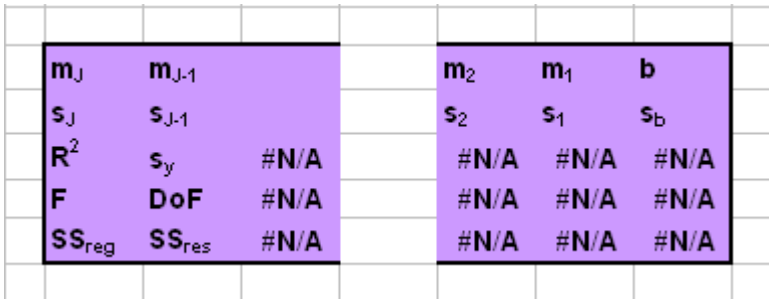


Рис. 8.2. Таблица вывода функция LINEST

$m_J, ..., m_2, m_1$ и b – оценки регрессионных коэффициентов;

$s_J, ..., s_2, s_1$ и s_b – стандартные ошибки для оценок регрессионных коэффициентов;

R^2 – коэффициент детерминации;

s_y – стандартная ошибка оценки y;

F – F-статистика;

DoF – число степеней свободы;

SS_{reg} – регрессионная сумма квадратов;

SS_{res} – остаточная сумма квадратов.

Примечания

- LINEST – это очень плохо сконструированная функция, очень неудобная в практическом применении;
- Примечания, представленные в описании функции TREND полностью применимы к функции LINEST.

Функция LINEST является функцией массива и ее ввод должен завершаться нажатием комбинации CTRL+SHIFT+ENTER.

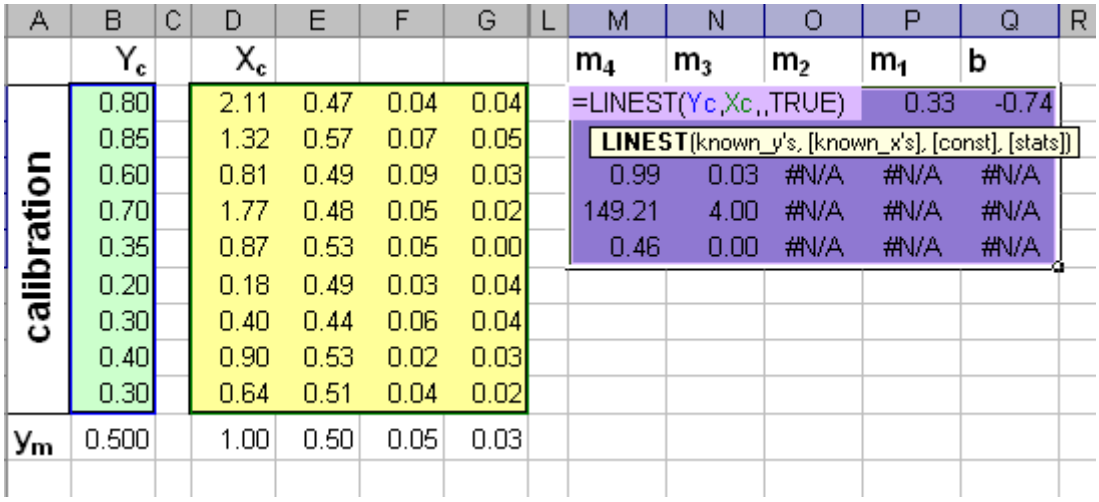


Рис. 8.3. Функция LINEST