Семинары по математической статистике (ФФ, 6-й семестр)

§1. Распределения случайных векторов

Определение. *Случайным вектором* $\vec{X} = (X_1, \ldots, X_n)$ называется такое отображение из пространства элементарных исходов Ω в n-мерное арифметическое простанство $\mathbf{R}^n$, что для каждого $\vec{t} = (t_1, \ldots, t_n) \in \mathbf{R}^n$ множество $\{\boldsymbol{\omega} \in \Omega : \ \vec{X}(\boldsymbol{\omega}) < \vec{t}\}$ является событием, то есть определена его вероятность. *Многомерной функцией распределения* случайного вектора $\vec{X}$ называется эта вероятность как функция векторной переменной $\vec{t}$:

$$F_{\vec{X}}(\vec{t}) = \mathbf{P}\{\, \boldsymbol{\omega} \in \Omega : \ \vec{X}(\boldsymbol{\omega}) < \vec{t}\} = \mathbf{P}\{\, \vec{X} < \vec{t}\}.$$

Неравенство $\vec{X} < \vec{t}$ понимается покоординатно, то есть означает систему неравенств $X_1 < t_1, \ldots, X_n < t_n$.

При $n = 1$ получаем обычные определения случайной величины и функции распределения.

Дискретным называется случайный вектор $\vec{X}$, принимающий конечное или счетное число значений $\vec{t}_1, \vec{t}_2, \ldots$. Его распределение задается таблицей распределения случайного вектора, т.е. набором вероятностей $\mathbf{P}\{\vec{X} = \vec{t}_j\}$. В двумерном случае (при $n = 2$) таблицу распределения дискретного случайного вектора $\vec{X} = (X, Y)$ удобно записывать в виде

$X \backslash Y$	b_1	b_2	$\ldots$
a_1	p_{11}	p_{12}	$\ldots$
a_2	p_{21}	p_{22}	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

Здесь $p_{ij} = \mathbf{P}\{X = a_i, \ Y = b_j\}$.

Свойства таблицы распределения двумерного дискретного вектора:

1) все a_i различны;

2) все b_j различны;

3) все p_{ij} неотрицательны;

4) сумма всех p_{ij} равна 1.

Таблица одномерного распределения случайной величины X получаются из таблицы двумерного распределения по формуле $\mathbf{P}\{X = a_i\} = \sum_j p_{ij}$.

Говорят, что случайный вектор $\vec{X}$ имеет *многомерное абсолютно непрерывное распределение*, если существует *многомерная плотность распределения вероятностей* $f_{\vec{X}}(\vec{t})$ такая, что для $B \subseteq \mathbf{R}^n$ выполнено равенство

$$\mathbf{P}\{\vec{X} \in B\} = \int\limits_B f_{\vec{X}}(\vec{t})d\vec{t}$$

(здесь и далее используется обозначение $d\vec{t} = dt_1 \dots dt_n$).

В частности, если $\vec{X}$ имеет абсолютно непрерывное распределение, то для любого $\vec{t} \in \mathbf{R}^n$ выполнено

$$F_{\vec{X}}(\vec{t}) = \int\limits_{\vec{u} \le \vec{t}} f_{\vec{X}}(\vec{u})d\vec{u}.$$

Свойства многомерной плотности распределения: $f_{\vec{X}}(\vec{t}) \ge 0$; $\int\limits_{\mathbf{R}^n} f_{\vec{X}}(\vec{t})d\vec{t} = 1$.

Одномерные плотности распределения компонент случайного вектора вычисляются интегрированием многомерной плотности распределения по всем значениям всех остальных компонент.

1.1. Найти вероятность p и одномерные таблицы распределения случайных величин X и Y. Найти $\mathbf{P}\{X = 0,\ Y > 1\}$, $F_{(X,Y)}(2,\ 2)$:

$X \backslash Y$	1	2	3
0	0,1	p	0
1	0	0	0,02
2	0,03	0	0

1.2. Найти вероятности p и q, если известно, что $\mathbf{P}\{X = -1\} = 0,3$. Найти таблицу распределения случайной величины Y и вероятность $\mathbf{P}\{X > 1,\ Y < 1\}$:

$X\backslash Y$	0	0,5	2
-1	0,1	p	0,05
3	0,3	0,4	q

1.3. Найти одномерные таблицы распределения случайных величин X и Y. Найти условное распределение X при условии $Y = -1$; при условии $Y = 2$. Найти условное распределение Y при условии $X = -2$; при условии $X = 1$.

$X\backslash Y$	-1	0	2
-2	0	0,5	0,1
1	0,1	0	0
3	0,3	0	0

1.4. Найти одномерные таблицы распределения случайных величин X и Y. Найти условное распределение X при условии $Y = 1$; при условии $Y > 1$. Найти условное распределение Y при условии $X = 0$.

$X\backslash Y$	-2	1	2	3	5
0	0,1	0,1	0,1	0,1	0,1
1	0,1	0,1	0,1	0,1	0,1

1.5. Найти константу A такую, чтобы функция $f(t_1, t_2) = At_1t_2\exp(-t_1^2 - t_2^2)$ при $t_1,\ t_2 \geq 0$, и равная нулю для всех остальных значений вектора (t_1, t_2), являлась двумерной плотностью распределения. Найти двумерную функцию распределения. Найти $F(0,0), F(0,-1), F(3,-2)$.

1.6. Плотность распределения $f(t_1, t_2)$ двумерного случайного вектора $(X_1,\ X_2)$ равна $\frac{1}{2}t_1t_2$ в треугольнике $\{0 \leq t_1 \leq 1,\ 0 \leq t_2 \leq 4t_1\}$, и нулю вне этого треугольника. Найти двумерную функцию распределения. Найти одномерные функции и плотности распределения. Найти $\mathbf{P}\{X_2 > 2\}, \mathbf{P}\{X_1 + X_2 < 1\}$.

1.7. Точку бросают наудачу в круг радиуса R с центром в начале координат. Пусть $(X,\ Y)$ — декартовы координаты точки. Найти плотность двумерного распределения, функции и плотности одномерных распределений. Найти $\mathbf{P}\{X < 0, Y < X\}$.

1.8*. Точку бросают наудачу в шар радиуса R с центром в начале координат. Пусть $(X,\ Y,\ Z)$ — декартовы координаты точки. Найти плотность трехмерного распределения, функции и плотности одномерных распределений.

1.9*. Пусть $(X_1,\ X_2,\ X_3)$ — координаты точки, брошенной наудачу в тетраэдр $\{t_1 \geq 0,\ \ t_2 \geq 0,\ \ t_3 \geq 0,\ \ t_1 + t_2 + t_3 \leq 2\}$. Найти плотность двумерного распределения $(X_1,\ X_2)$.

§2. Преобразования случайных векторов

Пусть g — функция n переменных, и $Y = g(\vec{X})$. Тогда случайная величина Y называется преобразованием случайного вектора $\vec{X}$ или функцией от случайного вектора $\vec{X}$.

Если $\vec{X}$ имеет дискретное распределение, то сначала находят всевозможные различные значения случайной величины Y, а потом вероятности, с которыми случайная величина принимает эти значения.

В абсолютно непрерывном случае из определения следует формула

$$F_{g(\vec{X})}(t) = \int\limits_{g(\vec{u})<t} f_{\vec{X}}(\vec{u})d\vec{u}.$$

Компоненты случайного вектора называются независимыми, если для любых подмножеств числовой прямой $B_1,\ \ldots,\ B_n$ выполнено равенство

$$\mathbf{P}\{X_1 \in B_1,\ \ldots,\ X_n \in B_n\} = \mathbf{P}\{X_1 \in B_1\} \cdot \ldots \cdot \mathbf{P}\{X_n \in B_n\}.$$

Распределение вектора с независимыми компонентами определяется распределениями компонент.

Пример. Случайные величины X_1, X_2 независимы и имеют абсолютно непрерывные распределения. Выразить плотность распределения их частного через их плотности распределения.

Пусть $Z = X_1/X_2$. Тогда

$$F_Z(u) = \mathbf{P}\{X_1/X_2 < u\}$$

$$= \int_0^\infty \mathbf{P}\{X_1 < uv,\ X_2 \in dv\} + \int_{-\infty}^0 \mathbf{P}\{X_1 > uv,\ X_2 \in dv\}$$

$$= \int_0^\infty F_{X_1}(uv) f_{X_2}(v) dv + \int_{-\infty}^0 (1 - F_{X_1}(uv)) f_{X_2}(v) dv.$$

Дифференцируя по u, получаем

$$f_Z(u) = \int_0^\infty v f_{X_1}(uv) f_{X_2}(v) dv - \int_{-\infty}^0 v f_{X_1}(uv) f_{X_2}(v) dv$$

$$= \int_{-\infty}^\infty |v| f_{X_1}(uv) f_{X_2}(v) dv.$$

2.1. Найти таблицы распределения случайных векторов $(X + Y,\ X - Y)$, $(\max(X,\ Y),\ \min(X,\ Y))$. Найти одномерные таблицы распределения случайных величин $X + Y$, $X - Y$, $\max(X,\ Y)$, $\min(X,\ Y)$, XY, Y/X.

$X\backslash Y$	-1	0	2
-2	0	0,5	0,1
1	0,1	0	0
3	0,3	0	0

2.2. Найти таблицы распределения случайных векторов $(X + Y,\ X - Y)$, $(\max(X,\ Y),\ \min(X,\ Y))$. Найти одномерные таблицы распределения случайных величин $X + Y$, $X - Y$, $\max(X,\ Y)$, $\min(X,\ Y)$, XY, Y/X.

$X\backslash Y$	-2	1	2	3	5
0	0,1	0,1	0,1	0,1	0,1
1	0,1	0,1	0,1	0,1	0,1

2.3. Случайные величины X_1, X_2, X_3 независимы, принимают значения 0 и 1 с равными вероятностями. Найти таблицы распределения случайных величин $X_1 + X_2 + X_3$, $X_1 + X_2 - X_3$, $2X_1 - X_2 - X_3$, $\max\{X_1,\ X_2, X_3\}$, $\min\{X_1,\ X_2, X_3\}$.

2.4. Случайные величины X_1, X_2 независимы, принимают значения -1, 0 и 1 с равными вероятностями. Найти таблицы распределения случайных величин $X_1 + X_2$, $X_1 - X_2$, $2X_1 - X_2$, $\max\{X_1,\ X_2\}$, $X_1 X_2^2$.

2.5. Случайная величина X имеет абсолютно непрерывное распределение с плотностью распределения f. Найти плотности распределения случайных величин $X+1$, $2-3X$, X^3, $\sqrt{|X|}$, $\sqrt{4+X^2}$.

2.6. Случайная величина X имеет абсолютно непрерывное распределение с плотностью распределения f. Найти плотности распределения случайных величин e^X, $1/X$, X^2, $\ln|X|$.

2.7. Найти плотность распределения суммы двух независимых случайных величин, имеющих показательное распределение с параметром α.

2.8. Найти плотность распределения суммы двух независимых случайных величин, имеющих равномерное распределение на отрезке $[0,\ 1]$.

2.9. Найти плотность распределения суммы двух независимых случайных величин, имеющих стандартное нормальное распределение.

2.10. Случайные величины X_1, X_2 независимы и имеют абсолютно непрерывные распределения. Выразить плотность распределения их произведения через их плотности распределения.

2.11. Случайные величины X_1, X_2 независимы и имеют абсолютно непрерывные распределения. Выразить плотность распределения их разности через их плотности распределения.

2.12. Случайные величины X_1, X_2 независимы и имеют абсолютно непрерывные распределения. Выразить плотность распределения случайной величины $X_1^3X_2^3$ через их плотности распределения.

2.13*. Случайные величины X_1, X_2 независимы и имеют равномерное распределение на отрезке $[0,\ 2]$. Найти плотность рапределения случайной величины $(X_1-X_2)^{-1}$.

2.14*. Случайные величины X_1, X_2 независимы и имеют показательное распределение с параметром 1. Найти плотность рапределения случайной величины X_1/X_2.

2.15. Случайные величины $X_1,\ldots,X_n$ независимы и имеют одно и то же абсолютно непрерывное распределение с плотно-

стью распределения f. Найти плотности распределения случайных величин $\max(X_1, \ldots, X_n)$ и $\min(X_1, \ldots, X_n)$.

§3. Моменты, ковариация, коэффициент корреляции

Если дискретный случайный вектор принимает значения $\vec{t}_1$, $\vec{t}_2$, ..., то его математическое ожидание — это вектор

$$\mathbf{E}\vec{X} = \sum_j \vec{t}_j \mathbf{P}\{\vec{X} = \vec{t}_j\}.$$

Если ряд

$$\sum_j |\vec{t}_j| \mathbf{P}\{\vec{X} = \vec{t}_j\}$$

расходится, то говорят, что математическое ожидание не существует. Здесь через $|\cdot|$ обозначена евклидова норма вектора.

Если $\vec{g}: \mathbf{R}^n \to \mathbf{R}^m$ — вектор-функция, $m \geq 1$, то

$$\mathbf{E}\vec{g}(\vec{X}) = \sum_j \vec{g}(\vec{t}_j) \mathbf{P}\{\vec{X} = \vec{t}_j\}.$$

В абсолютно непрерывном случае математическое ожидание определяется формулой

$$\mathbf{E}\vec{X} = \int\limits_{\mathbf{R}^n} \vec{t} f_{\vec{X}}(\vec{t}) d\vec{t}.$$

Математическое ожидание не существует, если

$$\int\limits_{\mathbf{R}^n} |\vec{t}| f_{\vec{X}}(\vec{t}) d\vec{t}$$

расходится.

Справедлива формула

$$\mathbf{E}\vec{g}(\vec{X}) = \int\limits_{\mathbf{R}^n} \vec{g}(\vec{t}) f_{\vec{X}}(\vec{t}) d\vec{t}.$$

Если компоненты вектора $\vec{g}$ записать в виде матрицы, соответствующие формулы дают математическое ожидание случайной матрицы.

Матрицей ковариаций случайного вектор-столбца $\vec{X}$ называется матрица

$$C(\vec{X}) = \mathbf{E}(\vec{X} - \mathbf{E}\vec{X})(\vec{X} - \mathbf{E}\vec{X})^T.$$

Матрица $C(\vec{X})$ симметрична и неотрицательно определена. Ее диагональные элементы — дисперсии компонент случайного вектора, а внедиагональные — ковариации соответствующих пар компонент.

Среднеквадратическим (стандартным) отклонением компоненты называется корень из ее дисперсии. Коэффициент корреляции двух компонент — это их ковариация, деленная на произведение стандартных отклонений.

3.1. Найти математические ожидания и дисперсии случайных величин X и Y. Найти ковариацию и коэффициент корреляции случайных величин X и Y.

$X\backslash Y$	-1	0	2
-2	0	0,5	0,1
1	0,1	0	0
3	0,3	0	0

3.2. Найти математические ожидания и дисперсии случайных величин X и Y. Найти ковариацию и коэффициент корреляции случайных величин X и Y. Найти математические ожидания и дисперсии случайных величин $X + Y$ и $2X - 3Y - 2$.

$X\backslash Y$	-2	1	2	3
0	0,2	0	0	0,2
1	0,1	0,2	0,2	0,1

3.3. Наудачу выбирают цифру от 0 до 9. Найти коэффициент корреляции индикаторов событий «цифра делится без остатка на 3» и «цифра делится без остатка на 5».

3.4. Из 20 студентов 5 написали на «отлично» первую контрольную, 4 — вторую контрольную, и 3 — обе контрольные.

Для выбранного наудачу студента найти коэффициент корреляции индикаторов событий «первая контрольная написана на отличную оценку» и «вторая контрольная написана на отличную оценку».

3.5. Плотность распределения $f(t_1, t_2)$ двумерного случайного вектора $(X_1,\ X_2)$ равна $\frac{1}{2}t_1t_2$ в треугольнике $\{0 \le t_1 \le 1,\ 0 \le t_2 \le 4t_1\}$, и нулю вне этого треугольника. Найти математические ожидания и дисперсии компонент случайного вектора. Найти их ковариацию и коэффициент корреляции.

3.6. Точку бросают наудачу в круг радиуса R с центром в начале координат. Пусть $(X,\ Y)$ — декартовы координаты точки. Найти математические ожидания и дисперсии координат точки. Найти их ковариацию и коэффициент корреляции.

3.7. Случайные величины X и Y имеют дисперсии σ_X^2, σ_Y^2 и коэффициент корреляции ρ. Найти такую константу c, чтобы X и $Y - cX$ были некоррелированными.

3.8*. Найти коэффициент корреляции случайных величин X и X^2, если X имеет показательное распределение.

§4. Матрица ковариаций. Многомерное нормальное распределение

Пусть случайный вектор $\vec{X}$ имеет многомерное стандартное нормальное распределение, то есть его многомерная плотность распределения вероятностей равна

$$f_{\vec{X}}(\vec{t}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\vec{t}^{\,T}\vec{t}\right),$$

где $\vec{t} = (t_1,\ \ldots,\ t_n)^T$, и $\vec{t}^{\,T}\vec{t} = t_1^2 + \ldots + t_n^2$.

Пусть вектор-столбец $\vec{Y}$ выражается через вектор-столбец $\vec{X}$ линейным образом:

$$\vec{Y} = \vec{a} + B\vec{X},$$

где $\vec{a}$ — неслучайный вектор-столбец, B — ненулевая квадратная матрица. Тогда говорят, что $\vec{Y}$ имеет многомерное нормальное распределение.

Нормальный вектор $\vec{Y}$ имеет вектор математического ожидания $\mathbf{E}\vec{Y} = \vec{a}$ и ковариационную матрицу $C = C(\vec{Y}) = BB^T$. Распределение нормального вектора $\vec{Y}$ полностью определяется математическим ожиданием и ковариационной матрицей. Матрица B определяется по данному многомерному нормальному распределению с точностью до ортогональной матрицы.

Многомерное нормальное распределение называется невырожденным, если матрица B невырождена, то есть $\det B \neq 0$, или, что эквивалентно, $\det C \neq 0$. В этом случае существует многомерная плотность распределения

$$f_{\vec{Y}}(\vec{t}) = \frac{1}{(2\pi)^{n/2}(\det C)^{1/2}} \exp\left(-\frac{1}{2}\vec{t}^{\,T} C^{-1}\vec{t}\right).$$

4.1. Найти вектор математического ожидания и матрицу ковариаций случайного вектора $(X,\ Y, X+Y)$:

$X\backslash Y$	-1	0	1
0	0,6	0	0
1	0	0	0,2
2	0	0,2	0

4.2. Найти вектор математического ожидания и матрицу ковариаций случайного вектора $(X,\ Y, XY)$. Найти математические ожидания и дисперсии случайных величин $X + Y$ и $XY - X - Y$.

$X\backslash Y$	0	2
0	0,1	0
1	0,7	0,2

4.3. Записать формулу плотности распределения 4-мерного стандартного нормального вектора. Найти плотность распределения суммы его компонент.

4.4. Найти вероятность того, что все компоненты стандартного нормального n-мерного вектора имеют один и тот же знак.

4.5. Пусть (X, Y, Z) — трехмерный стандартный нормальный вектор. С помощью таблицы нормального распределения найти приближенно $\mathbf{P}\{X > 1,96,\ Y < -2,33,\ Z < 0\}$.

4.6. Пусть (X, Y) — двумерный стандартный нормальный вектор. С помощью таблицы нормального распределения найти приближенно $\mathbf{P}\{X > -1,96,\ Y < 2,33\}$.

4.7. Пусть (X, Y) — двумерный стандартный нормальный вектор. Найти вероятность того, что $0 < X < Y$.

4.8. Пусть (X, Y) — двумерный стандартный нормальный вектор. Найти вероятность того, что $X < Y < X\sqrt{3}$.

4.9. Пусть

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 & -4 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

вектор $(X_1,\ X_2)$ имеет стандартное нормальное распределение.

Записать плотность распределения вектора $(Y_1,\ Y_2)$.

Записать в виде двукратного интеграла вероятность $\mathbf{P}\{Y_1 > -2,\ Y_1 + Y_2 < 2\}$.

4.10*. Пусть

$$Y_1 = X_1 + X_2, \quad Y_2 = X_1 - 2X_2,$$

вектор $(X_1,\ X_2)$ имеет нормальное распределение с нулевым вектором математического ожидания, единичными дисперсиями компонент и коэффициентом корреляции компонент, равным $1/2$. Записать плотность распределения вектора $(Y_1,\ Y_2)$. Записать в виде двукратного интеграла вероятность того, что обе компоненты вектора $(Y_1,\ Y_2)$ положительны.

§5. Характеристические и производящие функции

Если случайная величина X принимает только неотрицательные целые значения, то ее производящая функция $\mathbf{E}z^X = \sum_{k=0}^{\infty} p_k z^k$, где $p_k = \mathbf{P}\{X = k\}$.

Здесь z может быть комплексным числом, и математическое ожидание — это комплексное число, чьи действительная и мнимая части — математические ожидания действительной и мнимой частей комплексной случайной величины.

Для восстановления распределения неотрицательной целочисленной случайной величины достаточно рассмотреть производящую функцию для действительных z, так как разложение в ряд Тейлора единственно.

Для произвольной случайной величины X определение производящей функции $\mathbf{E}z^X$ потребовало бы выбора главной ветви функции $z^{X(\omega)}$ на каждом элементарном исходе ω. Но в рассмотрении всех комплексных z нет необходимости. Достаточно рассмотреть точки единичного круга $z = e^{it}$, где t принимает действительные значения. Получаем преобразование Фурье распределения случайной величины X, называемое характеристической функцией:

$$\varphi_X(t) = \mathbf{E}e^{itX}.$$

В дискретном случае

$$\varphi_X(t) = \sum_k e^{ita_k}\mathbf{P}\{X = a_k\}.$$

В абсолютно непрерывном случае

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itu} f_X(u)du.$$

Из теории преобразования Фурье известно, что распределение однозначно восстанавливается по характеристической функции.

Согласно теореме о непрерывном соответствии, сходимость последовательности характеристических функций $\varphi_{X_n}(t)$ в каждой точке $t \in \mathbf{R}$ к некоторой предельной характеристической функции $\varphi_X(t)$ эквивалентна сходимости по распределению последовательности случайных величин X_n к случайной величине X. Сходимость по распределению означает, что последовательность функций распределения случайных величин X_n сходится к предельной функции распределения во всех точках, за исключением точек разрыва этой предельной функции.

Если $\varphi_X(t) = e^{ita}$, то X имеет вырожденное распределение в точке a, и сходимость по распределению означает, что для любого $\varepsilon > 0$ имеет место сходимость

$$\mathbf{P}\{|X_n - a| \geq \varepsilon\} \to 0$$

при $n \to \infty$.

Если $\varphi_X(t) = \exp(-t^2/2)$, то X имеет стандартное нормальное распределение, и сходимость по распределению означает, что в любой точке u последовательность функций распределения $F_{X_n}(u)$ сходится к функции Лапласа

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp(-v^2/2)dv.$$

Характеристическая функция случайного вектора $\vec{X}$ — это функция векторной переменной $\vec{t}$:

$$\varphi_{\vec{X}}(\vec{t}) = \mathbf{E}e^{i\vec{t}^T\vec{X}} = \mathbf{E}\exp(i(t_1X_1 + \ldots + t_nX_n)).$$

Для характеристических функций случайных векторов также имеет место теорема о непрерывном соответствии.

5.1. Найти характеристическую и производящую функции случайной величины, принимающей значения 0, 1 и 2 с равными вероятностями.

5.2. Найти характеристическую и производящую функции бернуллиевского распределения.

5.3. Найти характеристическую и производящую функции биномиального распределения.

5.4. Найти характеристическую и производящую функции пуассоновского распределения.

5.5. Пусть X — неотрицательная целочисленная случайная величина. Выразить $\mathbf{E}X$ и $\mathbf{D}X$ через производные производящей функции.

5.6. Найти характеристическую функцию показательного распределения.

5.7. Найти характеристическую функцию гамма-распределения.

5.8. Найти характеристическую функцию квадрата стандартной нормальной случайной величины.

5.9. Найти характеристическую функцию суммы квадратов n независимых стандартных нормальных случайных величин.

5.10. По характеристическим функциям восстановить распределения: $\cos t$, $(1-4it)^{-1}$, $\exp(2it-2t^2)$.

5.11*. Найти представление характеристической функции двумерного нормального вектора через вектор математического ожидания и ковариационную матрицу.

5.12*. Найти плотность двумерного распределения, соответствующего характеристической функции $\exp\left(-2t_1^2+t_1t_2-2t_2^2\right)$.

§6. Предельные теоремы

Определение. Последовательность случайных величин $\{Y_n\}$ называется сходящейся с вероятностью единица к случайной величине Y, если

$$\mathbf{P}\{\omega:\ Y_n(\omega)\to Y(\omega)\}=\mathbf{P}\{Y_n\to Y\}=1.$$

Обозначение: $Y_n \xrightarrow{1} Y$.

Теорема (усиленный закон больших чисел, УЗБЧ). *Пусть случайные величины $X_1, X_2, \ldots$ независимы и одинаково распределены, причем $\mathbf{E}|X_1|<\infty$. Обозначим $a=\mathbf{E}X_1$, $S_n=\sum_{i=1}^n X_i$. Тогда при $n\to\infty$*

$$\frac{S_n}{n}\xrightarrow{1} a.$$

Центральная предельная теорема (ЦПТ). *Пусть $X_1, X_2\ldots$ — независимые одинаково распределенные случайные величины. Предположим, что $\mathbf{E}X_1^2<\infty$. Обозначим $S_n=$*

$X_1 + \ldots + X_n$, $a = \mathbf{E}X_1$, $\sigma^2 = \mathbf{D}X_1$, *и пусть* $\sigma^2 > 0$. *Тогда для любого* y

$$\mathbf{P}\left\{\frac{S_n - na}{\sigma\sqrt{n}} < y\right\} = F_{\frac{S_n - na}{\sigma\sqrt{n}}}(y) \to \Phi_{0,1}(y) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{y} e^{-t^2/2}\, dt$$

при $n \to \infty$.

Центральная предельная теорема для случайных векторов. *Пусть* $\vec{X}_1, \vec{X}_2 \ldots$ — *независимые одинаково распределенные случайные векторы с вектором математического ожидания* $\vec{a}$ *и ненулевой ковариационной матрицей* C. *Обозначим* $\vec{S}_n = \vec{X}_1 + \ldots + \vec{X}_n$. *Тогда для любого* $\vec{y}$

$$\mathbf{P}\left\{\frac{\vec{S}_n - n\vec{a}}{\sqrt{n}} < \vec{y}\right\} \to \mathbf{P}\{\vec{Z} < \vec{y}\}$$

при $n \to \infty$, *где* $\vec{Z}$ — *нормальный случайный вектор с нулевым математическим ожиданием и ковариационной матрицей* C.

Пример 6.1. К чему сходится с вероятностью единица при $n \to \infty$ последовательность

$$Y_n = \cos\frac{X_1 + \ldots + X_n}{n},$$

если $X_1, \ldots, X_n$ — независимые случайные величины, распределенные равномерно на $[0; \pi]$?

Решение. В силу закона больших чисел

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{1} \mathbf{E}X_1 = \frac{\pi}{2}.$$

Функция $\cos t$ непрерывна, поэтому

$$Y_n = \cos\frac{X_1 + \ldots + X_n}{n} \xrightarrow{1} \cos(\pi/2) = 0.$$

Пример 6.2. 1000 раз бросается игральная кость. Найти пределы, в которых с вероятностью 0,95 будет лежать сумма выпавших очков.

Решение. Обозначим через S_n сумму выпавших очков. S_n есть сумма независимых случайных величин, каждая из которых принимает значения от 1 до 6 с равными вероятностями. Нетрудно вычислить: $a = \mathbf{E}X_1 = 3,5$; $\mathbf{E}X_1^2 = 91/6$; $\sigma^2 = \mathbf{D}X_1 = 35/12$. В силу ЦПТ случайная величина $(S_n - 3500)/\sqrt{1000 \cdot 35/12}$ имеет почти стандартное нормальное распределение (число n велико!), поэтому

$$\mathbf{P}\left\{-1,96 < \frac{S_n - 3500}{\sqrt{1000 \cdot 35/12}} < 1,96\right\} \approx \frac{1}{\sqrt{2\pi}} \int\limits_{-1,96}^{1,96} e^{-t^2/2}\, dt = 0,95.$$

Последнее мы заранее находим из таблиц. Таким образом,

$$\mathbf{P}\left\{|S_n - 3500)| < 1,96\sqrt{1000 \cdot 35/12}\right\} \approx 0,95,$$

$$1,96\sqrt{1000 \cdot 35/12} \approx 106.$$

Итак, с вероятностью, близкой к 0,95, сумма выпавших очков лежит в пределах от 3394 до 3606.

6.1 Игрок в каждой игре (независимо от результатов других игр) выигрывает 80 рублей с вероятностью 0,1, проигрывает 20 рублей с вероятностью 0,9. Найти, к какой величине сходится средний выигрыш за n игр при $n \to \infty$.

6.2 Пусть $X_1,\ X_2,\ \ldots$ — случайные числа, то есть независимые случайные величины, имеющие равномерное распределение на отрезке от 0 до 1. Найти пределы п. н. следующих выражений при $n \to \infty$:

а) $\dfrac{{X_1}^2 + \ldots + {X_n}^2}{n}$; в) $\dfrac{1}{n}\left(\dfrac{1}{1+X_1} + \ldots + \dfrac{1}{1+X_n}\right)$;

б) $\dfrac{\sqrt{{X_1}^2 + \ldots + {X_n}^2}}{n}$; г) $\operatorname{arctg}\left(\dfrac{2}{n}\left(X_1 + \ldots + X_n\right)\right)$.

6.3. Случайные величины $X_1, X_2, \ldots$ независимы и одинаково распределены по закону Пуассона с параметром $\boldsymbol{\lambda}$. К чему сходится с вероятностью единица последовательность

$$\frac{X_1^2 + \ldots + X_n^2}{n} - \left(\frac{X_1 + \ldots + X_n}{n}\right)^2 \quad ?$$

6.4. Случайные величины $X_1,\ X_2,\ \ldots$ независимы и равномерно распределены на отрезке $[0,\ a]$. Доказать, что $Y_n \to a$ с вероятностью единица при $n \to \infty$ для последовательности случайных величин $Y_n = \max(X_1,\ \ldots,\ X_n)$ (указание: использовать тот факт, что для сходимости монотонной последовательности к константе a с вероятностью единица достаточно сходимости функций распределения во всех точках, отличных от точки a).

6.5 Какова вероятность того, что в 100 партиях одинаковых по силе противников один из них выиграет более 70 раз? Ничьих нет.

6.6. Вероятность выхода из строя за время T одного конденсатора равна 0,05. Определить вероятность того, что за время T из 100 конденсаторов выйдут из строя: а) не менее 5 конденсаторов; б) менее 13 конденсаторов.

6.7. Студент получает на экзамене 5 с вероятностью 0,2, 4 с вероятностью 0,4, 3 с вероятностью 0,3 и 2 с вероятностью 0,1. За время обучения он сдает 100 экзаменов. Найти пределы, в которых с вероятностью 0,95 лежит средний балл.

6.8. Урожайность куста картофеля задается следующим распределением:

Урожай в кг	0	1	1,5	2	2,5
Вероятность	0,1	0,2	0,2	0,3	0,2

На участке высажено 900 кустов. В каких пределах с вероятностью 0,95 будет находиться урожай? Какое наименьшее число

кустов нужно посадить, чтобы с вероятностью не менее 0,975 урожай был не менее тонны?

6.9*. Игральная кость подбрасывается до тех пор, пока общая сумма очков не превысит 700. Оценить вероятность того, что для этого потребуется более 210 бросаний.

6.10*. Пусть $X_1, X_2, \ldots$ — независимые одинаково распределенные случайные величины, $\mathbf{E}X_1 = 0, \quad \mathbf{D}X_1 < \infty$. Известно, что

$$\mathbf{P}\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}} \geq \ 1\right) \to \frac{1}{3}$$

при $n \to \infty$. Найти $\mathbf{D}X_1$.

6.11. Известно, что вероятность рождения мальчика приблизительно равна 0,515. Какова вероятность того, что среди 10 тыс. новорожденных окажется мальчиков не больше, чем девочек?

6.12. Для лица, дожившего до двадцатилетнего возраста, вероятность смерти на 21-м году жизни равна 0,006. Застрахована группа 10000 лиц 20-летнего возраста, причем каждый застрахованный внес 1200 рублей страховых взносов за год. В случае смерти застрахованного родственникам выплачивается 100000 рублей. Какова вероятность того, что:

а) к концу года страховое учреждение окажется в убытке;

б) его доход превысит 6000000 рублей?

Какой минимальный страховой взнос следует учредить, чтобы в тех же условиях с вероятностью 0,95 доход был не менее 4000000 рублей?

6.13. Суммируются 100 независимых одинаково распределенных векторов с нулевым вектором математического ожидания, равными единице дисперсиями компонент и коэффициентом корреляции компонент, равным -1/2. Записать в виде двойного интеграла приближенную вероятность того, что каждая из компонент суммы будет меньше 30.

6.14. На светофоре загорается красный, желтый или зеленый свет с равными вероятностями. Найти приближенно вероят-

ность того, что при 90 наблюдениях светофора студент менее 20 раз заставал зеленый свет. Записать в виде двойного интеграла приближенную вероятность того, что при 90 наблюдениях светофора студент более 40 раз заставал красный свет и более 30 раз желтый.

§7. Выборка. Оценивание параметров

Выборка и вариационный ряд

Основным объектом исследования в математической статистике является **выборка** $\vec{X} = (X_1, X_2, ..., X_n)$, то есть набор значений случайной величины X, полученных в результате n независимых воспроизведений эксперимента. Иначе говоря, выборка представляет собой случайный вектор, координаты которого — **элементы выборки** $X_1, X_2, ..., X_n$ — независимые случайные величины, имеющие общее распределение с функцией распределения $F(t)$. Будем говорить в этом случае, что имеется **случайная выборка** $\vec{X}$ из распределения F, и обозначать сокращенно: $\vec{X} \Subset F$. Число n называется **объемом выборки**. Конкретный набор числовых значений случайных величин $X_1, X_2, ..., X_n$, полученный в результате эксперимента, будем называть **реализацией** выборки и обозначать $\vec{x} = (x_1, x_2, ..., x_n)$.

Если элементы выборки $X_1, \ldots, X_n$ упорядочить по возрастанию, то получится новый набор случайных величин, называемый **вариационным рядом**:

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

Случайная величина $X_{(k)}$, $k = 1, \ldots, n$ называется **k-м членом вариационного ряда**, или **k-й порядковой статистикой**. В частности, $X_{(1)} = \min\{X_1, \ldots, X_n\}$, $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Эмпирическая функция распределения, гистограмма

Эмпирической функцией распределения $F_n^*(t)$ называется частота элементов выборки, меньших заданного t. Эмпирическая функция распределения, соответствующая выборке $\vec{X} = (X_1, X_2, ..., X_n)$, может быть построена по этой выборке с помощью любой из следующих формул:

$$F_n^*(t) = \frac{\{\text{количество } X_i : X_i < t\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(X_i < t),$$

где функция

$$\mathbf{I}(X_i < t) = \begin{cases} 1, & \text{если } X_i < t; \\ 0 & \text{иначе;} \end{cases}$$

— индикатор события $\{X_i < t\}$.

Заметим, что эмпирическая функция распределения, соответствующая случайной выборке $\vec{X}$, сама является случайной, поскольку определяется через элементы выборки $X_1,\ X_2, \ldots, X_n$, являющиеся случайными величинами. В то же время любая реализация $\vec{x} = (x_1, x_2, \ldots, x_n)$ выборки $\vec{X}$ порождает соответствующую реализацию эмпирической функции распределения (по той же формуле), которая является обычной (а не случайной) функцией распределения.

Эмпирическая функция распределения $F_n^*(t)$ является выборочным аналогом неизвестной теоретической функции распределения $F(t)$, ее называют также **оценкой** для $F(t)$. Выборочным аналогом для теоретической плотности распределения $f(t)$ является **гистограмма**, или **эмпирическая плотность распределения**, которая строится по выборке $\vec{X} = (X_1, X_2, ..., X_n)$ следующим образом.

Пусть $h > 0$ — произвольное число. Разобьем область значений изучаемой случайной величины (например, всю числовую ось) на промежутки $\Delta_k = [z_{k-1}, z_k)$ длины h и построим ступенчатую функцию $f_n^*(t)$, которая на каждом промежутке Δ_k

принимает постоянное значение, вычисляемое по любой из формул:

$$f_n^*(t) = \frac{\nu_k}{nh} = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{I}(X_i \in \Delta_k),\ t \in \Delta_k, \qquad (1)$$

где ν_k — число элементов выборки, попавших в промежуток Δ_k.

Иногда шаг гистограммы h выбирают следующим образом. Сначала рассчитывают число интервалов K по *формуле Стеджеса*

$$K = [\log_2 n] + 1. \qquad (2)$$

Здесь n — объем выборки, $[a]$ — целая часть числа a. Потом длина интервала рассчитывается по формуле

$$h = \frac{X_{(n)} - X_{(1)}}{K}.$$

При построении гистограммы последний промежуток выбирается замкнутым: $\Delta_K = [z_{K-1};\ z_K]$. Величину $X_{(n)} - X_{(1)} = \max\{X_i\} - \min\{X_i\}$ называют размахом выборки.

Выборочные моменты

По выборке $\vec{X} = (X_1, X_2, ..., X_n)$ можно построить эмпирические (выборочные) аналоги числовых характеристик распределения. Наиболее употребительными являются выборочное математическое ожидание, или **выборочное среднее**, $\overline{X}$, и **выборочная дисперсия** S^2:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2. \qquad (3)$$

Подобно выборочным среднему и дисперсии определяются выборочные моменты порядка k

$$\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

которые являются эмпирическими аналогами моментов $\mathbf{E}X_i^k$. Отметим, что

$$\mathbf{E}\overline{X^k} = \mathbf{E}X_i^k.$$

Приведенное соотношение означает, что математические ожидания эмпирических моментов совпадают с соответствующими теоретическими моментами. Это свойство называется **несмещенностью**. Эмпирические моменты являются **несмещенными оценками** для соответствующих теоретических.

Обобщая понятие выборочного момента, построим выборочное усреднение произвольной функции g:

$$\overline{g(X)} = \frac{1}{n}\sum_{i=1}^{n} g(X_i),$$

при этом также выполняется свойство

$$\mathbf{E}\overline{g(X)} = \mathbf{E}g(X_i).$$

Центральным выборочным моментом порядка k называется

$$\overline{(X - \overline{X})^k} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k.$$

Таким образом, второй центральный выборочный момент — это выборочная дисперсия S^2.

Центральные выборочные моменты являются смещенными оценками для своих теоретических аналогов $\mathbf{E}(X_i - \mathbf{E}X_i)^k$.

Корень из выборочной дисперсии $S = \sqrt{S^2}$ называется выборочным среднеквадратическим (стандартным) отклонением.

Отметим, что выборочная дисперсия вычисляется аналогично дисперсии.

$$S^2 = \overline{X^2} - (\overline{X})^2.$$

Несмещенная выборочная дисперсия — это статистика

$${S_0}^2 = \frac{n}{n-1}S^2.$$

Для нее выполнено свойство

$$\mathbf{E}{S_0}^2 = \mathbf{D}X_1.$$

Отметим, что корень из несмещенной выборочной дисперсии S_0 не является несмещенной оценкой для стандартного отклонения $\boldsymbol{\sigma}_X$, так как $\mathbf{E}\sqrt{Y} \neq \sqrt{\mathbf{E}Y}$.

Выборочной асимметрией называется

$$\widetilde{\mathbf{As}} = \frac{\overline{(X - \overline{X})^3}}{S^3},$$

выборочным эксцессом

$$\widetilde{\mathbf{Ex}} = \frac{\overline{(X - \overline{X})^4}}{S^4}.$$

Они являются выборочными аналогами асимметрии и эксцесса, определяемых формулами

$$\mathbf{As}X_1 = \frac{\mathbf{E}(X_1 - \mathbf{E}X_1)^3}{\boldsymbol{\sigma}^3}, \quad \mathbf{Ex}X_1 = \frac{\mathbf{E}(X_1 - \mathbf{E}X_1)^4}{\boldsymbol{\sigma}^4},$$

где $\boldsymbol{\sigma} = \sqrt{\mathbf{D}X_1}$ — теоретическое стандартное отклонение.

Асимметрия и эксцесс — безразмерные характеристики распределения.

Асимметрия характеризует скошенность распределения. Она равна 0 для симметричных распределений с конечным 3-м моментом.

Эксцесс характеризует вытянутость распределения. Он равен 3 для нормального распределения.

Статистики и оценки

Задача оценивания параметров возникает в ситуации, когда распределение F не является полностью неизвестным, а известен его математический вид $F = F(t, \boldsymbol{\theta})$, содержащий неизвестный параметр $\boldsymbol{\theta}$ (или несколько, тогда $\boldsymbol{\theta}$ — многомерный

параметр). Задача состоит в том, чтобы по выборке $\vec{X}$ вычислить приближенное значение $\theta^*(\vec{X})$ для неизвестного параметра, причем сделать это в том или ином смысле оптимальным образом. Это задача **точечного оценивания**.

Оценка $\tilde{\theta}$ называется **несмещенной** оценкой параметра θ, если для любого $\theta \in \Theta$ выполнено

$$\mathbf{E}\tilde{\theta} = \theta. \tag{4}$$

Договоримся указывать в обозначении статистики объем выборки, если это необходимо подчеркнуть: $\tilde{\theta} = \tilde{\theta}_n$.

Оценка $\tilde{\theta}_n$ называется **(сильно) состоятельной оценкой параметра** θ, если для любого $\theta \in \Theta$ при $n \to \infty$ имеет место сходимость с вероятностью единица:

$$\tilde{\theta}_n \xrightarrow{1} \theta, \tag{5}$$

то есть $\mathbf{P}\{\tilde{\theta}_n \to \theta\} = 1$.

Метод моментов (одномерный случай)

Оценкой метода моментов (ОММ) называется такое значение $\theta_g^* = \theta_g^*(\vec{X})$, при котором теоретическое среднее выборки $\vec{g}(X)$ совпадает с выборочным средним:

$$m_g(\theta_g^*) = \overline{g(X)},$$

то есть ОММ является решением уравнения относительно неизвестного θ_g^*.

Если при этом оказывается, что функция $m_g(\theta)$ непрерывна и строго монотонна, то для нее существует обратная m_g^{-1}, и ОММ имеет вид:

$$\theta_g^*(\vec{X}) = m_g^{-1}(\overline{g(X)}).$$

Отметим, что если функция $m_g(\theta) = \mathbf{E}g(X_1)$ непрерывна и строго монотонна, то оценка по методу моментов $\theta_g^*(\vec{X}) = m_g^{-1}(\overline{g(X)})$ сильно состоятельна.

Метод моментов (многомерный случай)

Пусть $\vec{X} \Subset \mathbf{F}_{\boldsymbol{\theta}}$, где параметр $\boldsymbol{\theta} \in \Theta$, подлежащий оцениванию, — многомерный. Рассмотрим для простоты двумерный случай, то есть $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Тогда для однозначного нахождения двух неизвестных $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ одного уравнения недостаточно. Оценкой метода моментов в этом случае называется решение $(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)$ системы уравнений вида:

$$\begin{cases} m_{g_1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \overline{g_1(X)}, \\ m_{g_2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \overline{g_2(X)}. \end{cases}$$

7.1. По данной реализации выборки $\vec{x}$ =(0; 0; 1; 1; 0; 0; 0; 0; 0; 1):

а) построить график реализации эмпирической функции распределения;

б) вычислить реализации выборочного среднего и выборочной дисперсии.

7.2 По реализации выборки 1; 0; 1; 1; 0; 1; 0; 0; 0; 1 вычислить реализации выборочного среднего, выборочной дисперсии, выборочного среднеквадратического отклонения, несмещенной выборочной дисперсии, выборочных асимметрии и эксцесса.

7.3. Измерен рост (в см) студентов одной учебной группы. Результаты измерений дали выборку (171; 186; 164; 190; 158; 181; 176; 180; 174; 157; 176; 169; 164; 186).

а) Построить реализацию гистограммы.

б) Вычислить реализации выборочного среднего, выборочной дисперсии и выборочного стандартного отклонения S. На одном графике с гистограммой построить график плотности нормального закона с параметрами $\overline{X}$, S^2.

7.4. Пусть $\vec{X} \Subset \Phi_{a,\boldsymbol{\sigma}^2}$. Вычислить $\mathbf{E}\overline{X}$, $\mathbf{D}\overline{X}$. Какое распределение имеет случайная величина $\overline{X}$?

7.5*. Пассажир маршрутного такси измерил 8 раз время ожидания такси и получил следующие результаты (в минутах): 8; 4; 5; 4; 2; 15; 1; 6. У него есть две гипотезы относительно

графика движения такси: либо график движения соблюдается, и время ожидания имеет равномерное распределение на отрезке $[0;\ \boldsymbol{\theta}]$, либо график движения не соблюдается, и время ожидания имеет показательное распределение с параметром $\boldsymbol{\lambda}$.

а) Вычислить реализации оценок параметров $\boldsymbol{\theta}$ и $\boldsymbol{\lambda}$, использовав оценки $\widetilde{\boldsymbol{\theta}}_2 = (n+1)X_{(n)}/n$ и $\widetilde{\boldsymbol{\lambda}}_2 = \frac{n-1}{n\overline{X}}$.

б) Построить на одном графике реализацию эмпирической функции распределения и теоретические функции распределения равномерного и показательного законов, в которые вместо неизвестных параметров подставлены реализации их оценок.

в) Построить на одном графике реализацию гистограммы и теоретические плотности распределения равномерного и показательного законов, в которые вместо неизвестных параметров подставлены реализации их оценок.

г) На основании проведенного исследования сделать вывод о том, какая из гипотез выглядит более соответствующей экспериментальным данным.

7.6*. Дана выборка $\vec{X} \Subset \Pi_{\boldsymbol{\lambda}}$, $\boldsymbol{\lambda} > 0$ — неизвестный параметр. Проверить, что статистики

$$T_1 = \overline{X},\quad T_2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}(X_i = k),\quad T_3 = \frac{X_1 + X_n}{2}$$

являются несмещенными оценками соответственно для $\boldsymbol{\lambda}$, $\frac{\boldsymbol{\lambda}^k}{k!}e^{-\boldsymbol{\lambda}}$ и $\boldsymbol{\lambda}$. Являются ли эти оценки состоятельными?

7.7. По выборке $(X_1,\ \ldots,\ X_n)$ из бернуллиевского распределения B_p с неизвестным параметром $p \in (0;\ 1)$ построить оценки параметра p:

а) по первому моменту;

б) по второму моменту;

в) по произвольному k-му моменту.

Можно ли отдать предпочтение какой-либо из построенных оценок? Исследовать их состоятельность и несмещенность.

7.8. По выборке $(X_1,\ \ldots,\ X_n)$ из биномиального распределения $B_{m,p}$ построить оценки методом моментов:

а) параметра p по первому и по второму моменту при известном $m > 0$;

б) параметров p и m.

Исследовать состоятельность построенных оценок.

7.9. Используя метод моментов, построить бесконечную последовательность различных оценок параметра θ равномерного распределения на отрезке $[0;\ \theta]$. Будут ли полученные оценки состоятельными?

7.10. С помощью метода моментов построить оценку параметра $\theta > 0$, если распределение выборки имеет плотность:

а) $\theta t^{\theta-1}$ при $t \in [0;\ 1]$; б) $2t/\theta^2$ при $t \in [0;\ \theta]$.

Исследовать полученные оценки на состоятельность.

7.11. Дана выборка из распределения с плотностью

$$f_\theta(t) = \begin{cases} 3t^2\theta^{-3}, & t \in [0;\ \theta]; \\ 0, & t \notin [0;\ \theta]. \end{cases}$$

Найти оценку параметра $\theta > 0$ методом моментов, исследовать ее на несмещенность и состоятельность.

7.12. Методом моментов найти оценку параметра $\alpha > 0$ по выборке из показательного распределения с плотностью $f_\alpha(t) = \alpha e^{-\alpha t},\ t > 0$. Будет ли оценка несмещенной и состоятельной?

7.13*. По выборке $(X_1, \ldots, X_n)$ методом моментов найти две различные оценки параметра $p \in (0, 1)$, если известно, что:

$P\{X_1 = 1\} = p/2;\ P\{X_1 = 2\} = p/2;\ P\{X_1 = 3\} = 1 - p.$

Будут ли полученные оценки несмещенными и состоятельными?

7.14*. При каких значениях параметра $\theta > 0$ распределения Парето с плотностью

$$f_\theta(t) = \begin{cases} \frac{\theta}{t^{\theta+1}}, & t \geq 1; \\ 0, & t < 1 \end{cases}$$

существует оценка параметра по первому моменту? Можно ли построить состоятельную оценку методом моментов в случае, когда оценки по первому моменту не существует?

7.15*. По выборке $(X_1, \ldots, X_n)$ из распределения Лапласа с плотностью $f_\lambda(t) = \frac{\lambda}{2}e^{-\lambda|t|}, \quad t \in \mathbf{R}$, построить оценку параметра $\lambda > 0$ методом моментов.

7.16. Пусть дана выборка из нормального распределения с параметрами α и σ^2. Используя метод моментов, построить оценки:

а) неизвестного математического ожидания α;

б) неизвестной дисперсии σ^2, если α известно;

в) неизвестной дисперсии σ^2, если α неизвестно.

Исследовать полученные оценки на несмещенность и состоятельность.

7.17*. Используя метод моментов, оценить параметр θ равномерного распределения на отрезке:

а) $[-\theta;\ \theta]$, $\theta > 0$; б) $[\theta;\ \theta + 1]$.

Исследовать полученные оценки на несмещенность и состоятельность.

§8. Оценки максимального правдоподобия

Пусть $\vec{X} \Subset F(t, \theta),\ \theta \in \Theta$. Предположим, что теоретическое распределение либо абсолютно непрерывно с плотностью $f(t, \theta) = f_{X_i}(t)$, либо дискретно, при этом для ряда распределения будем использовать то же обозначение: $f(t, \theta) = \mathbf{P}\{X_i = t\}$. **Функцией правдоподобия, соответствующей выборке** $\vec{X}$, называется функция

$$\Pi(\theta) = \Pi(\vec{X},\ \theta) = \prod_{i=1}^{n} f(X_i,\ \theta).$$

Оценкой максимального правдоподобия (ОМП) называется такое значение параметра $\theta = \hat{\theta}(\vec{X})$, при котором функция правдоподобия принимает наибольшее значение, то есть

$$\Pi(\vec{X},\ \hat{\theta}) = \max_{\theta \in \Theta} \Pi(\vec{X},\ \theta).$$

8.1. По выборке $(X_1,\ \ldots,\ X_n)$ из бернуллиевского распределения B_p с неизвестным параметром $p \in (0;\ 1)$ построить оценку параметра p методом максимального правдоподобия. (Указание: показать, что вероятность попадания в точку t для элементов выборки равна $f(t,\ p) = p^t(1-p)^{1-t}$, где t может принимать только два значения — 0 и 1). Исследовать состоятельность и несмещенность полученной оценки.

8.2. По выборке $(X_1,\ \ldots,\ X_n)$ из биномиального распределения $B_{m,p}$ построить оценку максимального правдоподобия параметра p при известном $m > 0$. Исследовать состоятельность и несмещенность оценки.

8.3. По выборке из показательного распределения E_α построить оценку максимального правдоподобия параметра $\alpha > 0$. Исследовать состоятельность оценки.

8.4. Построить оценку максимального правдоподобия по выборке из распределения Парето с плотностью

$$f_\theta(t) = \begin{cases} \frac{\theta}{t^{\theta+1}}, & t \geq 1; \\ 0, & t < 1. \end{cases}$$

Доказать состоятельность полученной оценки.

8.5*. С помощью метода максимального правдоподобия построить оценку параметра $\theta > 0$, если элементы выборки имеют плотность распределения:

а) $\theta t^{\theta-1}$ при $t \in [0;\ 1]$; б) $2t/\theta^2$ при $t \in [0;\ \theta]$.

Исследовать полученные оценки на состоятельность.

8.6*. Дана выборка из распределения с плотностью

$$f_\theta(t) = \begin{cases} 3t^2\theta^{-3}, & t \in [0;\ \theta]; \\ 0, & t \notin [0;\ \theta]. \end{cases}$$

Найти оценку параметра $\theta > 0$ методом максимального правдоподобия, исследовать ее на несмещенность и состоятельность.

8.7*. По выборке $(X_1, \ldots, X_n)$ методом максимального правдоподобия найти оценку параметра $p \in (0, 1)$, если известно, что $P\{X_1 = 1\} = p/2,\ P\{X_1 = 2\} = p/2,\ P\{X_1 = 3\} = 1 - p$.

Будет ли полученная оценка несмещенной и состоятельной?

8.8. Дана выборка из распределения с плотностью

$$f_\theta(t) = \begin{cases} e^{\theta - t}, & t \geq \theta; \\ 0, & t < \theta. \end{cases}$$

Найти оценку для θ:

а) методом моментов;

б) методом максимального правдоподобия.

Будут ли полученные оценки состоятельными? Вычислить смещения оценок и получить исправленные несмещенные оценки.

8.9*. По выборке $(X_1, \ldots, X_n)$ из распределения Лапласа с плотностью $f_\lambda(t) = \frac{\lambda}{2} e^{-\lambda|t|}, \quad t \in \mathbf{R}$, построить оценку параметра $\lambda > 0$ методом максимального правдоподобия.

8.10. Пусть дана выборка из нормального распределения с параметрами α и σ^2. Используя метод максимального правдоподобия, построить оценки:

а) неизвестного математического ожидания α;

б) неизвестной дисперсии σ^2, если α известно;

в) неизвестной дисперсии σ^2, если α неизвестно.

Исследовать полученные оценки на несмещенность и состоятельность.

8.11*. Используя метод максимального правдоподобия, оценить параметр θ равномерного распределения на отрезке:

а) $[-\theta;\ \theta]$, $\theta > 0$;

б) $[\theta;\ \theta + 1]$.

Исследовать полученные оценки на несмещенность и состоятельность.

§9. Сравнение оценок: среднеквадратический подход

Пусть $\vec{X} \Subset F(t, \theta), \quad \theta \in \Theta, \quad$ и $\widetilde{\theta} = \widetilde{\theta}(\vec{X})$ — какая-нибудь оценка параметра θ. Так как оценка является случайной величиной, то даже свойство несмещенности не гарантирует бли-

зость ее конкретной реализации $\widetilde{\theta}(\vec{x})$ к оцениваемому параметру. Если оценка является состоятельной, то такая близость гарантируется с заданной вероятностью, но только при достаточно больших объемах выборки n. При фиксированном объеме выборки наиболее распространенной «мерой близости» оценки к оцениваемому параметру является среднее значение квадрата отклонения $\mathbf{E}(\widetilde{\theta}-\theta)^2$.

Из двух оценок $\widetilde{\theta}_1$ считается *лучше*, чем $\widetilde{\theta}_2$, если при всех $\theta \in \Theta$ выполняется неравенство

$$\mathbf{E}(\widetilde{\theta}_1-\theta)^2 \le \mathbf{E}(\widetilde{\theta}_2-\theta)^2,$$

а хотя бы для одного θ неравенство является строгим.

Заметим, что $\mathbf{E}(\widetilde{\theta}-\theta)^2$ не меньше дисперсии оценки, и равенство достигается для несмещенных оценок:

$$\mathbf{E}(\widetilde{\theta}-\theta)^2 = \Big(\mathbf{E}(\widetilde{\theta}-\theta)\Big)^2 + \mathbf{D}(\widetilde{\theta}-\theta) = \Big(\mathbf{E}\widetilde{\theta}-\theta\Big)^2 + \mathbf{D}\widetilde{\theta} \ge \mathbf{D}\widetilde{\theta}.$$

Если $\widetilde{\theta}$ — несмещенная оценка параметра θ, то есть $\mathbf{E}\widetilde{\theta}=\theta$, то для нее:

$$\mathbf{E}(\widetilde{\theta}-\theta)^2 = \Big(\mathbf{E}\widetilde{\theta}-\theta\Big)^2 + \mathbf{D}\widetilde{\theta} = \mathbf{D}\widetilde{\theta}.$$

Отметим, что при среднеквадратическом подходе к сравнению оценок нельзя найти наилучшую в классе всех оценок (в частности, существуют несравнимые оценки). Доказательство этого факта основано на рассмотрении вырожденных оценок, равных константе независимо от значений выборки.

Для того, чтобы избежать необходимости сравнивать получаемые оценки с вырожденными оценками, нужно ограничить класс рассматриваемых оценок. Как правило, сравнивают только несмещенные оценки. Среди несмещенных оценок наилучшая оценка параметра для заданного параметрического семейства может существовать. Ее называют *эффективной* оценкой. Эффективная оценка имеет наименьшую дисперсию из всех несмещенных оценок.

Для однопараметрического семейства плотностей $f_\theta(y)$ информацией Фишера называется функция

$$I(\theta) = \mathbf{E}\left(\frac{\partial}{\partial\theta}\ln f_\theta(X_1)\right)^2.$$

Однопараметрическое семейство плотностей $f_\theta(y)$ будем называть *регулярным*, если информация Фишера хорошо определена в следующем смысле:

1) если для данного y логарифм плотности $\ln f_\theta(y)$ определен хотя бы для одного значения θ, то он непрерывно дифференцируем по параметру θ в области всех его возможных значений;

2) информация Фишера существует, положительна и непрерывна по θ.

Теорема (неравенство Рао—Крамера) Если однопараметрическое семейство плотностей регулярно, а $\widetilde{\theta}$ — несмещенная оценка его параметра, то

$$\mathbf{D}\widetilde{\theta} \geq \frac{1}{nI(\theta)}.$$

Из теоремы следует, что если для несмещенной оценки параметра регулярного семейства достигается равенство

$$\mathbf{D}\widetilde{\theta} = \frac{1}{nI(\theta)},$$

то оценка эффективна.

Многомерному параметру θ сопоставляется информационная матрица Фишера.

9.1. Имеется выборка четного объема n из распределения с конечной ненулевой дисперсией. По этой выборке построены 2 оценки математического ожидания: среднее по всей выборке и среднее по первой половине выборки. Сравнить их в среднеквадратическом смысле.

9.2. Пусть $\vec{X}$ — выборка из распределения с математическим ожиданием θ и конечной ненулевой дисперсией σ_θ^2. Выяснить, каковы должны быть константы $C_1, \ldots, C_n$, чтобы оценки вида $\widetilde{\theta} = C_1X_1 + C_2X_2 + \cdots + C_nX_n$ были несмещенными. Показать, что оценка $\theta_1^* = \overline{X}$ является наилучшей в среднеквадратическом в этом классе оценок.

9.3. Для выборок из следующих распределений найти оценку максимального правдоподобия $\widehat{\theta}$, проверить ее несмещенность и вычислить $\mathbf{E}(\widehat{\theta} - \theta)^2$:

1) распределение Бернулли с параметром p;

2) биномиальное распределение с параметрами 2, p;

3) геометрическое распределение с параметром $1/\theta$, $\theta > 1$ (напомним, что $\mathbf{P}_\theta\{X = k\} = \frac{1}{\theta}\left(1 - \frac{1}{\theta}\right)^{k-1}$, $k \geq 1$, $\mathbf{E}X_1 = \theta$, $\mathbf{D}X_1 = \theta(\theta - 1)$);

4) показательное распределение с параметром $1/\theta$, $\theta > 0$;

5) нормальное распределение с параметрами a, 1;

6) нормальное распределение с параметрами 0, σ^2.

9.4. Дана выборка $\vec{X} \Subset \mathrm{U}_{[0,\theta]}$; $\theta > 0$ — неизвестный параметр. Сравнить, какая из оценок для параметра θ лучше в среднеквадратическом смысле: $\theta_1^* = 2\overline{X}$, $\widetilde{\theta} = \frac{n+1}{n}X_{(n)}$.

9.5. Для распределений из задачи 9.3 проверить условие регулярности, вычислить информацию Фишера и исследовать эффективность полученных в задаче 9.3 оценок максимального правдоподобия.

9.6*. Дана выборка из распределения с плотностью

$$f_\theta(t) = \begin{cases} e^{\theta - t}, & t \geq \theta; \\ 0, & t < \theta. \end{cases}$$

Найти оценки для θ методом моментов и методом максимального правдоподобия. Сравить найденные оценки в среднеквадратическом.

9.7*. По выборке $(X_1, \ldots, X_n)$ из распределения Лапласа с плотностью $f_\lambda(t) = \frac{\lambda}{2}e^{-\lambda|t|}$, $t \in \mathbf{R}$, построить оценки параметра $\lambda > 0$ на основании второго момента и методом максимального

правдоподобия. Сравнить эти оценки в среднеквадратическом смысле.

§10. Оценивание параметров в задачах линейной регрессии

Наблюдается случайная величина y, значения которой зависят от случайного вектора **факторов регрессии** $\overrightarrow{x} = (x_1, ..., x_k)$. Введем в рассмотрение вектор-столбец неизвестных параметров регрессии $\overrightarrow{\theta} = (\theta_1, \ldots, \theta_k)^T$. Будем изучать линейную регрессию

$$Y_i = \sum_{j=1}^{k} x_{ij}\theta_j + \varepsilon_i, \quad i = 1, \ldots, n.$$

Предполагается, что случайные величины $\varepsilon_1, \ldots, \varepsilon_n$ некоррелированы и распределены с нулевым математическим ожиданием и конечной ненулевой дисперсией σ^2.

В матричном виде

$$\vec{Y} = X\vec{\theta} + \vec{\varepsilon}.$$

Сформулируем следующую теорему.

Теорема Гаусса—Маркова Пусть матрица X имеет ранг k. Тогда оценка $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)^T$, полученная по методу наименьших квадратов, которая минимизирует функцию

$$L(\vec{\theta}) = (\overrightarrow{Y} - X\overrightarrow{\theta})^T(\overrightarrow{Y} - X\overrightarrow{\theta}),$$

является несмещенной и имеет вид

$$\widehat{\theta} = (X^T X)^{-1} X^T \overrightarrow{Y}.$$

Ковариационная матрица оценки $\widehat{\theta}$ вычисляется по формуле

$$C(\widehat{\theta}) = \sigma^2 (X^T X)^{-1}.$$

Если $n > k$, то несмещенная оценка параметра $\boldsymbol{\sigma}^2$ равна

$$\widehat{\boldsymbol{\sigma}^2} = \frac{1}{n-k}||\overrightarrow{Y} - X\widehat{\boldsymbol{\theta}}||^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2.$$

Здесь $||\cdot||^2$ — скалярный квадрат вектора, $\widehat{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^T = X\widehat{\boldsymbol{\theta}}$.

Величина

$$\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{(n-k)\widehat{\boldsymbol{\sigma}^2}}{nS_Y^2}$$

— доля выборочной дисперсии, не объясненной регрессионной моделью.

Коэффициент детерминации

$$R^2 = 1 - \frac{(n-k)\widehat{\boldsymbol{\sigma}^2}}{nS_Y^2}$$

— это доля объясненной выборочной дисперсии.

10.1. Пусть $Y_i = x_i + \boldsymbol{\theta} + \varepsilon_i$, $i = 1, \ldots, n$. Здесь x_i, $\boldsymbol{\theta} \in \mathbf{R}$. Найти оценку для $\boldsymbol{\theta}$ по методу наименьших квадратов. Найти оценку дисперсии регрессионных ошибок $\boldsymbol{\sigma}^2$.

10.2. Пусть $Y_i = \boldsymbol{\theta}x_i + \varepsilon_i$, $i = 1, \ldots, n$. Здесь x_i, $\boldsymbol{\theta} \in \mathbf{R}$. Выяснить, для каких значений x_i выполнены предположения теоремы Гаусса—Маркова. Найти оценку для $\boldsymbol{\theta}$ по методу наименьших квадратов. Найти оценку дисперсии регрессионных ошибок $\boldsymbol{\sigma}^2$.

10.3. Концентрация лекарства $Y > 0$ в крови пациента обратно пропорциональна массе тела $x > 0$. Найти оценку коэффициента пропорциональности для следующих моделей:

1) $Y_i = \boldsymbol{\theta}/x_i + \varepsilon_i$;

2) $\ln Y_i = \ln(\boldsymbol{\theta}/x_i) + \varepsilon_i$;

$i = 1, \ldots, n$. Найти оценку параметра $\boldsymbol{\theta}$ в каждой модели. Найти дисперсию оценки в первой модели и дисперсию логарифма оценки во второй модели.

10.4. Для регрессионной модели $Y_i = a + bx_i + \varepsilon_i$, $i = 1, \ldots, n$, найти оценки параметров a, b по методу наименьших квадратов. Найти ковариационную матрицу оценок. Найти оценку дисперсии регрессионных ошибок $\boldsymbol{\sigma}^2$.

10.5. По реализации двумерной выборки $x_1 = 1$, $Y_1 = 0$, $x_2 = 2$, $Y_2 = 2,5$, $x_3 = 3$, $Y_3 = 0,5$, найти реализации оценок параметров модели из задачи 11.4. Вычислить реализацию коэффициента детерминации.

10.6*. Для регрессионной модели $Y_i = a_1 \cos x_i + b_1 \sin x_i + \varepsilon_i$, $i = 1, \ldots, n$, найти оценки параметров a_1, b_1 по методу наименьших квадратов. Рассмотреть случай $x_i = \boldsymbol{\pi} i/2$, $n = 4$. Найти ковариационную матрицу оценок.

§11. Интервальное оценивание

Пусть имеется выборка объема n из распределения, известного с точностью до параметра: $\vec{X} \Subset F(t, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. *Доверительным интервалом с уровнем доверия* $1 - \varepsilon$ для неизвестного параметра $\boldsymbol{\theta}$ называют случайный интервал $(\boldsymbol{\theta}_-;\ \boldsymbol{\theta}_+) \subset \Theta$, построенный по выборке, который накрывает неизвестное значение параметра с вероятностью, равной $1 - \varepsilon$, или по крайней мере стремящейся к $1 - \varepsilon$ с ростом объема выборки, то есть

$$\mathbf{P}\{\boldsymbol{\theta} \in (\boldsymbol{\theta}_-;\ \boldsymbol{\theta}_+)\} \to 1 - \varepsilon$$

при $n \to \infty$.

В случае, когда вместо сходимости выполняется точное равенство, доверительный интервал называется *точным*.

$\boldsymbol{\theta}_-, \boldsymbol{\theta}_+$ — это оценки параметра $\boldsymbol{\theta}$, называемые *нижней и верхней доверительными границами*. Число $1 - \varepsilon \in (0;\ 1)$ — уровень доверия, или доверительная вероятность, — выбирается заранее и отражает «степень готовности мириться с возможностью ошибки». Чем менее мы готовы мириться с возможной ошибкой, тем меньшее (более близкое к нулю) значение ε должны устанавливать.

Асимптотические доверительные интервалы

Если распределение не является нормальным, точный доверительный интервал, как правило, не удается построить. Поэтому строят асимптотический доверительный интервал, применяя центральную предельную теорему, которая утверждает, что для всех $t_1, t_2 \in \mathbf{R}$ $(t_1 < t_2)$ выполнено:

$$\lim_{n\to\infty} \mathbf{P}\left\{ t_1 \le \frac{n\overline{g(X)} - n\mathbf{E}g(X_1)}{\sqrt{n\mathbf{D}g(X_1)}} < t_2 \right\} = \Phi(t_2) - \Phi(t_1),$$

то есть центрированные и нормированные суммы случайных величин $n\overline{g(X)} = g(X_1) + \ldots + g(X_n)$ сходятся по распределению к случайной величине, имеющей стандартное нормальное распределение.

Здесь предполагается, что $0 < \mathbf{D}g(X_1) < \infty$.

Если выбрать $t_2 = -t_1 = A$, $g(x) = x$ и принять доверительный уровень равным $1 - \varepsilon$, то

$$\lim_{n\to\infty} \mathbf{P}\left\{ -A \le \frac{n\overline{X} - n\mathbf{E}X_1}{\sqrt{n\mathbf{D}X_1}} < A \right\} = \Phi(A) - \Phi(-A) = 2\Phi(A) - 1 = 1 - \varepsilon,$$

откуда получаем:

$$\Phi(A) = 1 - \varepsilon/2.$$

По заданному ε можно найти A с помощью таблиц нормального распределения или программных приложений. Отметим следующее свойство сходимости по распределению: если Y_n сходится по распределению к Y, а Z_n сходится к 1 с вероятностью единица, то их произведение $Y_n Z_n$ сходится по распределению к Y. Обозначим $\boldsymbol{\sigma} = \sqrt{\mathbf{D}X_1}$ и выберем

$$Y_n = \frac{n\overline{X} - n\mathbf{E}X_1}{\boldsymbol{\sigma}\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - \mathbf{E}X_1)}{\boldsymbol{\sigma}}, \qquad Z_n = \frac{\boldsymbol{\sigma}}{S}.$$

Вспомним, что $S = \sqrt{\overline{X^2} - (\overline{X})^2} \to \boldsymbol{\sigma}$ с вероятностью 1, и, следовательно, $Z_n \to 1$ с вероятностью 1. Итак,

$$Y_n Z_n = \frac{\sqrt{n}(\overline{X} - \mathbf{E}X_1)}{\boldsymbol{\sigma}} \cdot \frac{\boldsymbol{\sigma}}{S} = \frac{\sqrt{n}(\overline{X} - \mathbf{E}X_1)}{S}$$

сходится по распределению к стандартной нормальной случайной величине, то есть

$$\lim_{n\to\infty} \mathbf{P}\left\{-A \le \frac{\sqrt{n}(\overline{X} - \mathbf{E}X_1)}{S} < A\right\} = \Phi(A) - \Phi(-A) = 2\Phi(A) - 1 = 1 - \varepsilon.$$

Чтобы для неизвестного параметра θ найти двусторонний доверительный интервал асимптотического уровня $1 - \varepsilon$, нужно для исследуемого однопараметрического семейства распределений найти зависимость $\mathbf{E}X_1 = a(\theta)$ и решить относительно параметра θ двойное неравенство:

$$-A \le \frac{\sqrt{n}(\overline{X} - a(\theta))}{S} < A.$$

Для этого нужно, чтобы функция $a(\theta)$ была непрерывной и строго монотонной. Получившиеся границы дверительного интервала будем обозначать через θ_- и θ_+.

Распределения, связанные с нормальным

При построении доверительных интервалов для параметров нормального распределения мы будем использовать два специальных распределения, связанных с нормальным: распределение хи-квадрат и распределение Стьюдента. Название «распределение Стьюдента» связано с именем английского статистика К.Госсета, который подписывал свои работы псевдонимом «Стьюдент».

Случайная величина Z_n имеет *распределение хи-квадрат с n степенями свободы*, если

$$Z_n = {X_1}^2 + \ldots + {X_n}^2;$$

где X_1, $\ldots$, X_n — независимые случайные величины со стандартным нормальным распределением.

Отметим, что «число степеней свободы» — это просто традиционное название для параметра n распределения хи-квадрат.

Параметр n — положительное целое число. В частности, при $n = 1$ получаем квадрат одной случайной величины со стандартным нормальным распределением: $Z_1 = X^2$, где $X \in N_{0,\,1}$.

Будем использовать следующее обозначение: $Z_n \in \boldsymbol{\chi}^2_n$.

Отметим следующие свойства распределения хи-квадрат.

Пусть $Z_n \in \boldsymbol{\chi}^2_n$. Тогда:

1) $\mathbf{E}Z_n = n$;

2) $Z_n/n \to 1$ с вероятностью единица при $n \to \infty$.

Случайная величина Y_n имеет *распределение Стьюдента с n степенями свободы*, если

$$Y_n = \frac{X}{\sqrt{Z_n/n}},$$

где случайные величины X и Z_n независимы, причем X имеет стандартное нормальное распределение, а Z_n имеет распределение хи-квадрат с n степенями свободы. Здесь, как и у распределения хи-квадрат, n — это просто положительный целый параметр.

Будем использовать следующее обозначение: $Y_n \in T_n$.

Отметим следующие свойства распределения Стьюдента.

Пусть $Y_n \in T_n$. Тогда:

1) для любого t выполнено $\mathbf{P}\{Y_n < -t\} = \mathbf{P}\{Y_n > t\}$, то есть распределение Стьюдента симметрично;

2) $Y_n \to X$ с вероятностью единица при $n \to \infty$, где X имеет стандартное нормальное распределение.

Точные доверительные интервалы

Наиболее распространенной ситуацией, когда возможно построение точных доверительных интервалов, является случай нормального распределения: $\vec{X} \in \Phi_{a,\boldsymbol{\sigma}^2}$, когда хотя бы один из его параметров неизвестен. В этом случае известно совместное распределение наиболее употребительных оценок $\overline{X}$ и S^2 параметров a и $\boldsymbol{\sigma}^2$, с помощью которого и строятся соответству-

ющие доверительные интервалы. Основные результаты содержатся в следующей теореме.

Теорема Фишера. Пусть $\vec{X} \in \Phi_{a,\sigma^2}$. Тогда верны следующие 4 факта.

1) $\dfrac{\sqrt{n}(\overline{X}-a)}{\sigma} \in \Phi_{0,1}$.

2) $\dfrac{\sum_{i=1}^{n}(X_i-a)^2}{\sigma^2} \in \chi^2_n$.

3) $\dfrac{nS^2}{\sigma^2} \in \chi^2_{n-1}$.

4) $\dfrac{\sqrt{n-1}\,\big(\overline{X}-a\big)}{S} \in T_{n-1}$.

11.1. Пусть элементы выборки $\vec{X}$ имеют плотность распределения

$$f(t) = \frac{1}{\pi(1+(t-\theta)^2)}, \quad t \in \mathbf{R}.$$

Здесь θ — неизвестный параметр, $\theta \in \mathbf{R}$. Построить точный доверительный интервал для параметра θ по одному наблюдению ($n = 1$).

11.2. $\vec{X} \in B_p$, $0 < p < 1$. Построить асимптотический доверительный интервал для параметра p.

11.3. Дана выборка из геометрического распределения с параметром p, $0 < p < 1$. Построить асимптотический доверительный интервал для параметра p.

11.4. По выборке из распределения Пуассона с параметром $\lambda > 0$ построить асимптотический доверительный интервал для параметра λ.

11.5*. Дана выборка из распределения с плотностью $e^{-|t-a|}/2$, $a \in \mathbf{R}$. Построить асимптотический доверительный интервал для параметра a.

11.6*. Пусть $\vec{X} \in U_{[0;\,\theta]}$, где $\theta > 0$. С помощью статистик

$\overline{X}$ и $\overline{X^2}$ построить асимптотические доверительные интервалы (соответственно (θ_1^-, θ_1^+) и (θ_2^-, θ_2^+)) уровня $1-\varepsilon$ и показать, что случайный интервал (θ_2^-, θ_2^+) асимптотически короче соответствующего (θ_1^-, θ_1^+).

11.7. Известно, что измерения величины a независимы, имеют нормальное распределение с математическим ожиданием a (то есть отсутствует систематическая погрешность) и стандартным отклонением 10 мм. Результаты 4 измерений дали среднее значение 512 мм. Найти доверительный интервал для параметра a уровня 0,95; уровня 0,998.

11.8. Известно, что измерения величины a независимы, имеют нормальное распределение с математическим ожиданием a (то есть отсутствует систематическая погрешность) и стандартным отклонением σ. Результаты 100 измерений эталонной длины 1 м дали выборочное среднее 1,01 м и выборочный второй момент 1,04 м^2. Найти доверительный интервал для стандартного отклонения уровня 0,9; уровня 0,99.

11.9. По выборке объема 25 из нормального распределения подсчитаны выборочное среднее 2,1 и выборочный второй момент 4,42. Построить точные доверительные интервалы уровня 0,95 для параметров нормального распределения.

§12. Статистические гипотезы и критерии

Пусть $\vec{X} = (X_1, X_2, ..., X_n)$ — выборка, $\vec{X} \Subset \mathbf{F}$, где $\mathbf{F}$ — полностью или частично неизвестное распределение отдельного наблюдения X_i.

Статистической гипотезой будем называть всякое утверждение о виде или свойствах неизвестного распределения $\mathbf{F}$.

Гипотеза называется *простой*, если она однозначно определяет распределение $\mathbf{F}$, в противном случае гипотеза называется *сложной*.

Мы будем рассматривать ситуацию, когда гипотез всего две. Одну из них называют *основной*, а другую — *альтернативной*,

обозначая соответственно H_0 и H_1.

Статистическим критерием называют всякое правило, позволяющее на основании наблюдаемого выборочного вектора $\vec{X}$ принять одну из гипотез: основную или альтернативную.

При применении статистического критерия могут возникнуть ошибки двух родов. Ошибка первого рода состоит в том, что отвергается верная нулевая гипотеза. Ошибка второго рода — отвергается верная первая гипотеза. Вообще ошибка i-го рода состоит в том, что статистический критерий отвергает верную $(i-1)$-ю гипотезу.

принимаемая гипотеза	верна гипотеза H_0	верна гипотеза H_1
H_0	нет ошибки	ошибка 2-го рода
H_1	ошибка 1-го рода	нет ошибки

Критерий характеризуется вероятностями ошибок:

$$\alpha_1 = \mathbf{P}_{H_0}(H_0 \text{ отвергается}); \qquad \alpha_2 = \mathbf{P}_{H_1}(H_1 \text{ отвергается}).$$

Здесь нижний индекс у символа вероятности указывает, при выполнении какой гипотезы подсчитывается вероятность.

Критерии согласия

Удобно представлять статистический критерий как функцию $\delta(\vec{X})$ от выборочного вектора, принимающую два значения: H_0 и H_1. Наиболее общий подход для построения статистических критериев состоит в следующем.

Пусть $T = T(\vec{X})$ — некоторая статистика, характеризующая отклонение эмпирических данных, представленных выборкой,

от теоретических, соответствующих проверяемой гипотезе H_0. Если распределение статистики $T(\vec{X})$ известно (точно или хотя бы приближенно), то для любого $\alpha > 0$ можно найти такое множество T_α значений T, для которого будет выполнено неравенство

$$\mathbf{P}(T \in \mathrm{T}_\alpha / \mathrm{H}_0) \leq \alpha.$$

Пусть $\alpha > 0$ настолько мало, что событие, имеющее вероятность, не превосходящую α, может считаться практически невозможным. Тогда статистический критерий можно задать следующим образом:

$$\delta(\vec{X}) = \begin{cases} \mathrm{H}_1, & \text{если } T(\vec{X}) \in \mathrm{T}_\alpha; \\ \mathrm{H}_0, & \text{если } T(\vec{X}) \notin \mathrm{T}_\alpha. \end{cases}$$

Число $\alpha > 0$, которое фигурирует в формулах, называется *уровнем критерия*, или *уровнем значимости*, статистика $T(\vec{X})$ называется *статистикой критерия*, а множество T_α — *критическим множеством*.

Достигаемый уровень значимости

От статистики $T = T(\vec{X})$ требуют следующих свойств:

1) при выполнении гипотезы H_0 статистика T имеет известное распределение или, по крайней мере, сходится по распределению к некоторой случайной величине J с известным распределением;

2) при выполнении гипотезы H_1 статистика T сходится почти наверное к бесконечности с ростом объема выборки.

Для того, чтобы получить критерий уровня α, задают критическое множество в виде

$$\mathrm{T}_\alpha = \{T \geq C\},$$

где C — константа, определяемая условием

$$\mathbf{P}\{J \geq C\} = \alpha,$$

то есть $F_J(C) = 1 - \alpha$.

Ясно, что при таком выборе константы C вероятность ошибки нулевого рода α_0 либо равна уровню критерия α (в случае, когда статистика T при верной нулевой гипотезе распределена в точности как J), либо, по крайней мере, сходится к α с ростом объема выборки.

Сходимость статистики T почти наверное к бесконечности при выполненной первой гипотезе гарантирует *состоятельность* критерия, то есть сходимость вероятности ошибки первого рода α_1 к нулю с ростом объема выборки.

Для каждой конкретной выборки $\vec{X}$ можно найти предельное значение уровня $\alpha^* = \alpha^*(\vec{X})$, при котором гипотеза H_0 еще может быть принята. Такое значение называется *реально достигаемым уровнем значимости*, или просто *достигаемым уровнем значимости*. Достигаемый уровень значимости α^* имеет смысл вероятности получить худшее согласие с проверяемой гипотезой, чем реально полученное, если гипотеза H_0 верна. Поэтому чем меньше α^*, тем более это говорит против гипотезы H_0.

Достигаемый уровень значимости вычисляется с помощью распределения статистики J:

$$\alpha^* = \mathbf{P}\{J \geq T(\vec{X})\} = 1 - F_J(T(\vec{X})).$$

В терминах достигаемого уровня значимости критическая область имеет вид

$$\mathrm{T}_\alpha = \{\alpha^* \leq \alpha\},$$

то есть нулевая гипотеза отвергается на уровне α в случае, когда $\alpha^* \leq \alpha$.

Каждый критерий согласия использует свою статистику, предназначенную для различения нулевой гипотезы и альтернативы и обладающую нужными свойствами: сходимостью к фиксированному распределению при выполнении нулевой гипотезы и сходимостью почти наверное к бесконечности при ее невыполнении.

В качестве важных примеров критериев согласия рассмотрим критерии Колмогорова и хи-квадрат Пирсона.

Критерии согласия Колмогорова и χ^2 Пирсона

Рассмотрим выборку $\vec{X} \Subset \mathrm{F}$ объема n с неизвестной функцией распределения F и простую гипотезу $\mathrm{H}_0 : \mathrm{F} = \mathrm{F}_0$. Альтернативной для H_0 является сложная гипотеза $\mathrm{H}_1 : \mathrm{F} \neq \mathrm{F}_0$.

Критерий Колмогорова применяется в случае, когда функция распределения $\mathrm{F}_0(t)$ непрерывна. Рассматривается следующее расстояние между эмпирической и теоретической функциями распределения:

$$D_n = D(F_n^*, F_0) = \sup_{-\infty<t<\infty} |F_n^*(t) - F_0(t)| = \max_{-\infty<t<\infty} |F_n^*(t) - F_0(t)|.$$

В качестве статистики критерия Колмогорова выбирается это расстояние, умноженное на $\sqrt{n}$, где n — объем выборки:

$$T_n = \sqrt{n} D_n = \sqrt{n} \max_{-\infty<t<\infty} |F_n^*(t) - F_0(t)|.$$

А.Н.Колмогоров доказал следующие свойства статистики T_n:

1) если гипотеза H_0 верна, то T_n с ростом n сходится к случайной величине J с функцией распределения, называемой функцией распределения Колмогорова:

$$F_J(t) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 t^2};$$

2) если гипотеза H_0 неверна, то T_n сходится почти наверное к $+\infty$ при $n \to \infty$. Таким образом, достигаемый уровень значимости критерия Колмогорова равен:

$$\alpha^* = 1 - F_J(T_n) = 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 T_n^2} = 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 n D_n^2}. \tag{6}$$

Отметим, что для расчетов по этой формуле нужно брать не всю бесконечную сумму, а только несколько слагаемых, при этом ошибка вычислений не превосходит последнего отброшенного слагаемого. Критерий Колмогорова отвергает гипотезу H_0 на уровне α, если $\alpha^* \leq \alpha$.

Для практического вычисления статистики $D_n = D_n(\vec{X})$ можно использовать следующую формулу:

$$D_n(\vec{X}) = \max_{1 \leq i \leq n} \max \left(\left| F(X_{(i)}) - \frac{i}{n} \right|; \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right).$$

Здесь $X_{(i)}$ — это элементы *вариационного ряда*, то есть для этих вычислений выборку следует предварительно *упорядочить по возрастанию*.

Если гипотетическая функция распределения $\mathrm{F}_0(x)$ не является непрерывной, то критерий Колмогорова неприменим. В этом случае можно воспользоваться χ^2*—критерием Пирсона*. Статистика критерия Пирсона строится после предварительного «группирования» выборочных данных. Для этого все множество S возможных значений случайных величин X_i разбивается на конечное число непересекающихся частей:

$$S = S_1 \cup S_2 \cup \cdots \cup S_r, \quad S_i \cap S_j = \emptyset, i \neq j.$$

Обозначим ν_j — число элементов выборки $\vec{X}$, попавших в множество S_j, а p_j — вероятность попадания случайной величины X_i в множество S_j, вычисленная с помощью гипотетической функции распределения $\mathrm{F} = \mathrm{F}_0$. Тогда в качестве статистики критерия χ^2 рассматривают следующую предложенную Пирсоном меру отклонения эмпирического распределения от предполагаемого теоретического:

$$\chi^2(\vec{X}) = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j}.$$

Справедлива следующая теорема, позволяющая находить распределение статистики χ^2 при больших значениях n, а стало быть, и строить статистический критерий.

Если гипотеза H_0 однозначно фиксирует вероятности $p_1,\ p_2,\ \ldots,\ p_r$, где $p_j = \mathbf{P}(X_i \in S_j)$, то при выполнении этой гипотезы статистика $\chi^2(\vec{X})$ слабо сходится к распределению χ^2_{r-1}:

$$\chi^2 \Longrightarrow \chi^2_{r-1}, \quad n \to \infty.$$

При невыполнении нулевой гипотезы статистика $\chi^2(\vec{X})$ сходится почти наверное к $+\infty$.

Для построения критерия, основанного на статистике χ^2, используем распределение χ^2_{r-1}, и по найденному значению $\chi^2(\vec{X})$ отыскиваем достигаемый уровень значимости:

$$\alpha^* = 1 - F_{\chi^2_{r-1}}(\chi^2(\vec{X}))$$

по таблице распределения хи-квадрат или с помощью математических пакетов. В пакете Microsoft Excel достигаемый уровень значимости вычисляется формулой

=ХИ2РАСП(ячейка;r-1)

(в качестве ячейки надо подставить адрес ячейки, в которой вычислена статистика хи-квадрат, а $r-1$ — число степеней свободы).

Тогда критерий Пирсона имеет следующий вид:

$$\mathrm{H}_0 \Leftrightarrow \ \alpha^* > \alpha.$$

Заметим, что для практического применения рекомендуется разбиение производить таким образом, чтобы выполнялось условие $np_j \geq 5$. При нарушении этого условия нужно объединить соседние множества S_j. Вероятности p_j надо выбирать по возможности равными.

Критерий хи-квадрат часто используют для проверки сложных гипотез о принадлежности распределения к некоторому параметрическому семейству (например, к нормальному). При этом вместо известных вероятностей p_j подставляют их оценки p_j^*, полученные путем оценивания неизвестных параметров

распределения. Важно понимать, что в этом случае предельное распределение статистики $\chi^2(\vec{X})$ уже не будет распределением χ^2_{r-1}, а будет близко к распределению χ^2_{r-1-s}, где s — число оцениваемых параметров ($s = 2$ для нормального распределения). Более точно, предельная функция распределения заключена между функциями распределения χ^2_{r-1-s} и χ^2_{r-1}.

Достигаемый уровень значимости α^* удовлетворяет неравенству:

$$1 - F_{\chi^2_{r-1-s}}(\chi^2(\vec{X})) \leq \alpha^* \leq 1 - F_{\chi^2_{r-1}}(\chi^2(\vec{X})),$$

где s — число оцениваемых параметров.

Для того, чтобы получить в точности распределение хи-квадрат с $r-1-s$ степенями свободы, следует оценивать неизвестные параметры методом максимального правдоподобия по *группированной* выборке, но это приводит, как правило, к сложным вычислительным процедурам.

12.1. Крупная партия товаров может содержать долю дефектных изделий. Поставщик полагает, что эта доля составляет 3%, а покупатель — 10%. Условия поставки: если при проверке 20 случайным образом отобранных товаров обнаружено не более одного дефектного, то партия принимается на условиях поставщика, в противном случае — на условиях покупателя. Требуется определить:
1) каковы статистические гипотезы, статистика критерия, область ее значений, критическая область;
2) какое распределение имеет статистика критерия, в чем состоят ошибки первого и второго рода и каковы их вероятности.

12.2. Имеется выборка объема 1 из нормального распределения $\Phi_{a,1}$. Проверяются простые гипотезы $H_0: a = 0$, $H_1: a = 1$. Используется следующий критерий (при заданной постоянной c):

$$H_0 \Leftrightarrow X_1 \leq c.$$

Вычислить, в зависимости от c, вероятности ошибок первого и второго рода.

12.3. Используя конструкции доверительного интервала, построить критерий точного уровня ε для проверки гипотезы $\mathrm{H} : \boldsymbol{\theta} = 1$, если:
а) $\vec{X} \in \Phi_{\boldsymbol{\theta},1}$;
б) $\vec{X} \in \Phi_{1,\boldsymbol{\theta}}$.

12.4. Построить критерий, обладающий нулевыми вероятностями ошибок, для проверки гипотез $\mathrm{H}_0 : \vec{X} \in \Phi_{0,1}$ против $\mathrm{H}_1 : \vec{X} \in \Pi_{\boldsymbol{\lambda}}$.

12.5. Пусть $\vec{X} \in \Phi_{a,1}$. Для проверки гипотез $\mathrm{H}_0 : a = 0$ против $\mathrm{H}_1 : a = 1$ используется следующий критерий: H_0 принимается, если $X_{(n)} < 3$, и отвергается в противном случае. Найти вероятности ошибок.

12.6*. Используя конструкции доверительного интервала, построить критерий асимптотического уровня ε для проверки гипотезы $\mathrm{H} : \boldsymbol{\theta} = 1$, если а) $\vec{X} \in \mathrm{E}_{\boldsymbol{\theta}}$; б) $\vec{X} \in \mathrm{B}_{\boldsymbol{\theta}/2}$; в) $\vec{X} \in \Pi_{\boldsymbol{\theta}}$.

12.7. Вычислить значение статистики Колмогорова по реализации выборки (1,1; 0,4; 0,2; 3,2), если основная гипотеза состоит в том, что распределение элементов выборки — равномерное на $[0,\ 4]$.

12.8. Вычислить достигнутый уровнень значимости критерия Колмогорова, если объем выборки равен 100, а $\sup_{-\infty<t<\infty} |F_n^*(t) - F_0(t)| = 0,2$.

12.9. При 4040 бросаниях монеты Бюффон получил $\boldsymbol{\nu}_1 = 2048$ выпадений герба и $\boldsymbol{\nu}_2 = n - \boldsymbol{\nu}_1 = 1992$ выпадений решетки. Согласуется ли это с гипотезой о том, что монета правильная, при уровне значимости 0,1? С каким предельным уровнем значимости может быть принята эта гипотеза?

12.10. При $n = 4000$ независимых испытаний события $A1$, $A2$, $A3$, составляющие полную группу, осуществились соответственно 1905, 1015 и 1080 раз. Проверить, согласуются ли эти данные при уровне значимости 0,05 с гипотезой H_0: $p_1 = 1/2$, $p_2 = p_3 = 1/4$, где $p_j = \mathbf{P}(A_j)$. Найти достигнутый уровень значимости.

12.11. В экспериментах с селекцией гороха Мендель на-

блюдал частоты различных видов семян, полученных при скрещивании растений с круглыми желтыми семенами и растений с морщинистыми зелеными семенами. Эти данные и значения теоретических вероятностей по теории наследственности приведены в следующей таблице:

Семена	Частота	Вероятность
Круглые и желтые	315	9/16
Морщинистые и желтые	101	3/16
Круглые и зеленые	108	3/16
Морщинистые и зеленые	32	1/16
Σ	n=556	1

Следует проверить гипотезу H_0 о согласовании частотных данных с теоретическими вероятностями (на уровне значимости 0,1) и найти достигнутый уровень значимости.

12.12. В таблице приведены числа m_i участков равной площади 0,25 км^2 южной части Лондона, на каждый из которых приходилось по i попаданий самолетов-снарядов во время второй мировой войны. Проверить согласие опытных данных с законом распределения Пуассона, приняв за уровень значимости $\alpha = 0,05$:

i	0	1	2	3	4	5 и более	Итого
m_i	229	211	93	35	7	1	$\Sigma m_i = 576$

§13. Статистические критерии для нескольких выборок

Проверка однородности двух выборок

Пусть $\vec{X}, \vec{Y}$ — независимые выборки объемов n и m соответственно. Гипотеза однородности утверждает, что эти выборки из одного и того же распределения.

Если распределение предполагается непрерывным, то применим критерий Колмогорова—Смирнова: вычислим статистику

$$d_{n,m} = \sqrt{\frac{mn}{n+m}} \sup_{t \in \mathbf{R}} |F^*_{1,n}(t) - F^*_{2,m}(t)|,$$

где $F^*_{1,n}(t)$, $F^*_{2,m}(t)$ — эмпирические функции распределения, построенные по выборкам $\vec{X}$ и $\vec{Y}$ соответственно. Если выполнена гипотеза однородности, то распределение статистики $d_{n,m}$ не зависит от конкретного распределения элементов выборки. Для больших n, m оно близко к распределению Колмогорова.

Если предполагается, что выборки из нормального распределения, то можно последовательно применить критерий Фишера для проверки равенства дисперсий и критерий Стьюдента для проверки равенства математических ожиданий.

При выполнении гипотезы равенства дисперсий нормальных выборок статистика

$$\frac{nS_x^2}{mS_y^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sum_{i=1}^m (Y_i - \overline{X})^2}$$

имеет распределение Фишера с $(n-1, m-1)$ степенями свободы.

Если равны дисперсии и математические ожидания нормальных выборок, то статистика

$$t_{n,m} = t_{n,m}(\vec{X}, \vec{Y}) = \sqrt{\frac{mn}{m+n}} \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{nS_x^2 + mS_y^2}{n+m-2}}}$$

имеет распределение Стьюдента с $n+m-2$ степенями свободы.

Проверка независимости

Если $(X_1, Y_1), \ldots, (X_n, Y_n)$ — выборка из двумерного нормального распределения, то гипотезу о независимости компонент выборки можно проверить с помощью выборочного коэффициента корреляции

$$\widehat{r}_n = \frac{\overline{XY} - \overline{X}\ \overline{Y}}{S_x S_y},$$

где $S_x^2 = \overline{X^2} - \overline{X}^2$, $S_y^2 = \overline{Y^2} - \overline{Y}^2$.

Если верна гипотеза о независимости, то выборочный коэффициент корреляции имеет плотность распределения

$$f_{\widehat{r}_n}(t) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-2}{2}\right)}(1-t^2)^{\frac{n-4}{2}},$$

$t \in (-1,\ 1)$, $n > 2$, а случайная величина

$$\widehat{r}_n\sqrt{\frac{n-2}{1-\widehat{r}_n^2}}$$

имеет распределение Стьюдента с $n-2$ степенями свободы.

13.1. Пусть $\vec{X}, \vec{Y}$ — независимые выборки объема 2 из непрерывного распределения. Составить таблицу распределения случайной величины $d_{2,2}(\vec{X}, \vec{Y})$.

13.2. По следующим реализациям выборок вычислить реализацию статистики $d_{n,m}$ Колмогорова—Смирнова:

$\vec{X}$ = (1,2, 0,4, -0,2, 0,9), $\vec{Y}$ = (0,2, -0,5, 1, -0,9, 0,3, 0,5).

13.3. По реализациям независимых выборок $\vec{X}, \vec{Y}$ объемов 40 и 50 вычислено значение $\sup_{t\in\mathbf{R}} |F^*_{1,n}(t) - F^*_{2,m}(t)| = 0,1$. Найти достигнутый уровень значимости гипотезы об однородности. Сделать вывод о том, принимается ли эта гипотеза на уровне 0,05.

13.4. По реализациям независимых выборок $\vec{X}, \vec{Y}$ объемов 20 и 30 из нормального распределения вычислены значения статистик $S_x^2 = 15$ и $S_y^2 = 10$. Найти реально достигнутый уровень значимости гипотезы о равенстве дисперсий против двусторонней альтернативы, а также против каждой из односторонних альтернатив.

13.5. Пусть в условиях предыдущей задачи предполагается равенство дисперсий, и известны значения $\overline{X} = 2, \overline{Y} = 12$. Найти реально достигнутый уровень значимости гипотезы о равенстве математических ожиданий против двусторонней альтернативы, а также против каждой из односторонних альтернатив.

13.6. Пусть $(X_1, Y_1), \ldots, (X_4, Y_4)$ — выборка объема 4 из двумерного нормального распределения. Найти, при каких значениях выборочного коэффициента корреляции гипотеза о независимости компонент отвергается на уровне 0,1 при двусторонней альтернативе. Вычислить значение выборочного коэффициента корреляции по реализации двумерной выборки (1, 2), (2, 3), (-1, 0), (0, 0).

13.7*. Пусть $(X_1, Y_1), \ldots, (X_{100}, Y_{100})$ — реализация выборки объема 100 из двумерного нормального распределения. Значение выборочного коэффициента корреляции $\widehat{r}_n$ равно –0,25. Найти реально достигнутый уровень значимости гипотезы о независимости компонент против двусторонней альтернативы, а также против каждой из односторонних альтернатив.

Таблица нормального распределения

Значения функции $\Phi(t) = \frac{1}{\sqrt{2\pi}} \cdot \int\limits_{-\infty}^{t} e^{-\frac{u^2}{2}} du$ и функции $\overline{\Phi}(t) = \Phi(-t) = 1 - \Phi(t)$.

t	$\Phi(-t)$	$\Phi(t)$
4,75	0,000001	0,999999
4,26	0,00001	0,99999
3,72	0,0001	0,9999
3,09	0,001	0,999
2,58	0,005	0,995
2,33	0,01	0,99
2,05	0,02	0,98
1,96	0,025	0,975
1,88	0,03	0,97
1,75	0,04	0,96
1,64	0,05	0,95
1,28	0,1	0,9
0,84	0,2	0,8
0,52	0,3	0,7
0,25	0,4	0,6
0,00	0,5	0,5

Для $|t| > 4,75$ можно использовать аппроксимацию $\overline{\Phi}(t) \sim \frac{e^{-t^2/2}}{t\sqrt{2\pi}}$.