МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ПУТЕЙ СООБЩЕНИЯ (МИИТ)

Кафедра “Прикладная математика–2”

А.С.МИЛЕВСКИЙ

ВЫСШАЯ МАТЕМАТИКА

ЧАСТЬ 4.
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Конспект лекций

МОСКВА – 2008

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ПУТЕЙ СООБЩЕНИЯ (МИИТ)

Кафедра “Прикладная математика–2”

А.С.МИЛЕВСКИЙ

ВЫСШАЯ МАТЕМАТИКА

ЧАСТЬ 4.
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Конспект лекций

Рекомендовано редакционно-
издательским советом
университета в качестве
конспектов лекций для
студентов ИЭФ и ИУИТ

МОСКВА – 2008

УДК-51
М-60
Милевский А.С. Высшая математика. Ч.4.
Математическая статистика: Конспект лекций. –
М.: МИИТ, 2008. – 44 с.

Конспект лекций предназначен для студентов, изучающих курс математического анализа в институтах ИЭФ и ИУИТ. Включает в себя материал по математической статистике.

Рецензенты:

Фролов Е.Б., д.т.н., профессор МГТУ Станкин,

Деснянский В.Н., к.ф.-м.н., заведующий кафедрой “Вычислительная математика” МИИТ.

Предмет математической статистики

Математическая статистика изучает математические методы обработки и анализа результатов статистических наблюдений.

Математическую модель изучаемой ситуации можно описать следующим образом.
Имеется случайная величина X, распределенная по неизвестному нам закону, значение которой измеряется в эксперименте. Повторим эксперимент n раз, предполагая, что условия его проведения не изменяются. Такой n-кратный составной эксперимент связан с n-мерной СВ (X_1, X_2,..., X_n), где X_j - СВ, соответствующая j-му эксперименту.

Ясно, что все X_j *независимы и одинаково распределены*; их общий закон распределения, совпадающий с законом распределения СВ X, называется *законом распределения генеральной совокупности*.

Вектор (X_1, X_2,..., X_n) называется *выборочным вектором*. Числа (x_1, x_2,..., x_n), получаемые в результате измерений, называют *реализацией выборочного вектора* или просто *выборкой*. Число **n** называется *объемом выборки*, а разность между наибольшим и наименьшим элементами – *ее размахом.*

Аналогично определяется *двумерная выборка,* отвечающая ситуации, когда в каждом эксперименте измеряются две случайные величины, X и Y.

Основная задача математической статистики – по выборке восстановить закон распределения случайной величины X как можно точнее.

Пример

Монету бросили 1000 раз, и выпало только 300 гербов. Можно ли считать, что для этой монеты вероятность выпадения герба меньше 1/2, или это получилось случайно?

Здесь X – случайная величина, распределённая по закону (0-герб, 1-решка):
Требуется **проверить гипотезу p<1/2**.

X	0	1
P	p	1-p

Способы описания выборки

1. Статистический ряд (таблица частот). Группированный статистический ряд.

X	X_1	X_2	…	X_k
n	n_1	n_2	…	n_k

Таблица частот.
Первая строчка содержит все *различные* значения из выборки, а вторая - сколько раз они встречаются.
Сумма чисел в нижней строке равна объему выборки n.

Группированная таблица частот.
Множество значений СВ разбито на r интервалов.
Вторая строка содержит количества элементов выборки, попавших в соответ-ствующий интервал. Если элемент попадает на границу двух соседних интервалов, то он учитывается в левом.

Δ	$Δ_1$	$Δ_2$	…	$Δ_r$
n	n_1	n_2	…	n_r

Пример 1.

Дана выборка

2,3,2,3,5,2,5,5,5,3,6,2,2,3,3

Заполнить таблицу частот

X_i	2	3	5	6
n_i	5	5	4	1

Пример 2.

Дана выборка

2,7,24,11,21,48,10,3,9,3,3,4,2,30,35,21, 11,14,21,5

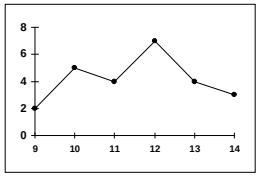
Заполнить группированную таблицу частот, разбив на 5 интервалов

Δ_i	0-10	10-20	20-30	30-40	40-50
n_i	10	3	5	1	1

Решение.

Интервалы можно выбирать одинаковой длины, а можно разной (в зависимости от выборки). Выберем одинаковой. Размах выборки = 48-2=46, так что длина интервала должна быть больше 46/5=9,25. Возьмём 10. Начало первого интервала должно быть не больше $x_{min}=2$. Возьмём 0.

2. Полигон частот.



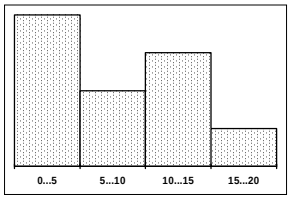
Полигон частот строится по таблице частот и представляет собой ломаную, соединяющую точки с координатами

(x_i,n_i) (i =1 ... k)

Задача.

По выборке из примера 1 построить полигон частот

3. Гистограмма.



Гистограмма строится по группированной таблице частот и представляет собой набор прямоугольников, основания которых совпадают с интервалами из таблицы, а высоты равны

$$h_i = \frac{n_i}{n \cdot l_i}$$

где l_i - длина i-го интервала.

Таким образом, общая площадь всех прямоугольников **равна 1**

Площадь части гистограммы над произвольным отрезком оси OX приближенно равна вероятности попадания в соответствующий интервал

Задача.

По выборке из примера 2 построить гистограмму

3. Эмпирическая функция распределения.

Эмпирическая функция распределения (ЭФР) F*(x) - это построенная по выборке приближенная функция распределения изучаемой случайной величины. Она определяется формулой

где в числителе стоит количество элементов выборки (с учетом кратности), меньших **x**

$$F^*(x) = \frac{\sum_{x_i < x} n_i}{n}$$

При большом объеме выборки ЭФР *близка* к *функции распределения генеральной совокупности*:

Теорема 1 (Гливенко). Пусть $F_n^*(x)$ - ЭФР, построенная по выборке объема n. Тогда для любого **x,**

то есть

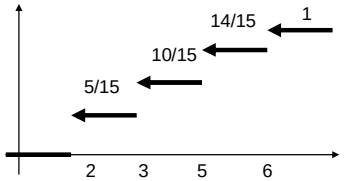
$$F_n^*(x) \xrightarrow[n\to\infty]{} F(x)$$

$$\lim_{n\to\infty} P\left(\left|F_n^*(x) - F(x)\right| > \varepsilon\right) = 0 \text{ для любого } \varepsilon$$

Пример 1.

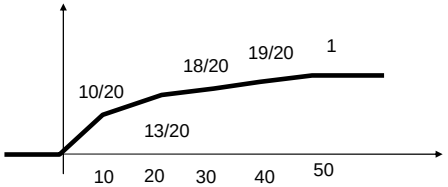
Для выборки из примера 1 (негруппированной)

X_i	2	3	5	6
n_i	5	5	4	1



Пример 2.

Для выборки из примера 1 (группированной)



Δ_i	0-10	10-20	20-30	30-40	40-50
n_i	10	3	5	1	1

Основные законы распределения из теории вероятностей

1) Закон Пуассона.

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2\ldots$$
$$MX=\lambda,\quad DX=\lambda$$

Число λ>0 – параметр закона Пуассона.

2) Биномиальный закон

$$P(X=k)=C_N^k p^k q^{N-k},\quad k=0,1,\ldots,N$$
$$MX=Np,\quad DX=Npq$$

Число N – количество испытаний,
p - вероятность успеха в одном испытании,
q=1-p.
X – количество успехов

3) Геометрический закон

$$P(X=k)=q^{k-1}\cdot p,\quad k=1,2,\ldots$$
$$MX=\frac{1}{p},\quad DX=\frac{q}{p^2}$$

Число N – количество испытаний, p - вероятность успеха в одном испытании, q=1-p.
X – количество испытаний до первого успеха включительно.

4) Равномерный закон на отрезке [A,B]

$$P(a\le X<b)=\frac{b-a}{B-A}\ при\ A\le a\le b\le B.$$

$$f(x)=\begin{cases}0, & x<A\\ \frac{1}{B-A}, & A\le x\le B\\ 0, & x>B\end{cases}$$

$$MX=\frac{A+B}{2},$$
$$DX=\frac{(B-A)^2}{12}.$$

5) Показательный закон

$$f(x)=\begin{cases}0, & x<0\\ \lambda e^{-\lambda x}, & x\ge 0,\end{cases}\quad \lambda>0.$$

$$P(a\le X<b)=e^{-\lambda a}-e^{-\lambda b}\ при\ 0\le a\le b$$

$$MX=\frac{1}{\lambda}.$$
$$DX=\frac{1}{\lambda^2}.$$

6) Нормальный закон

$$P(a\le X<b)=\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right)$$

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

$$MX=m.$$
$$DX=\sigma^2.$$

Точечные оценки

Часто закон распределения генеральной совокупности можно считать известным с точностью до одного или нескольких параметров. Возникает задача по выборке восстановить значения этих параметров по возможности точнее.

Точечной оценкой для неизвестного параметра Θ *называется приближенное значение этого параметра, найденное по выборке* $x=(x_1, x_2,..., x_n)$:

$$\tilde{\Theta} = \tilde{\Theta}(x_1, x_2,..., x_n).$$

Для неизвестного параметра можно придумать сколько угодно оценок, хороших и плохих. “Качество” оценки характеризуют понятия

Несмещённость

Оценка $\tilde{\Theta} = \tilde{\Theta}(x_1, x_2,..., x_n)$

называется несмещённой оценкой неизвестного параметра Θ, если

$$M\tilde{\Theta}(X_1, X_2,..., X_n) = \Theta$$

Другими словами, *среднее значение несмещённой оценки равно истинному значению параметра*.

Состоятельность

Оценка $\tilde{\Theta} = \tilde{\Theta}(x_1, x_2,..., x_n)$

называется состоятельной оценкой неизвестного параметра Θ, если

$$\tilde{\Theta}_n(X_1, X_2,..., X_n) \underset{n\to+\infty}{\overset{P}{\to}} \Theta$$

Другими словами, *с ростом объема выборки значение состоятельной оценки стремится к истинному значению параметра*.

Теорема. Обозначим $M_n = M\tilde{\Theta}_n(X_1,..., X_n)$, $D_n = D\tilde{\Theta}_n(X_1,..., X_n)$

Тогда если $\lim\limits_{n\to+\infty} D_n = 0$, $\lim\limits_{n\to+\infty} M_n = \Theta$,

то оценка является состоятельной.

Эффективность

Говорят, что оценка $\tilde{\Theta}_{(1)}$ *эффективнее*, чем оценка $\tilde{\Theta}_{(2)}$,

если ее дисперсия меньше.

Некоторые точечные оценки

Оценка для MX называется "***выборочное среднее***", обозначается $\overline{x}$

и вычисляется как среднее арифметическое чисел - элементов выборки (с учетом кратностей):

$$\overline{x} = \frac{x_1 n_1 + \ldots + x_k n_k}{n}$$

Эта оценка всегда *несмещённая* и *состоятельная*

Оценка для дисперсии DX называется "*выборочная дисперсия*", обозначается D*x и вычисляется по формуле:

$$D_x^* = \overline{(x-\overline{x})^2} = \overline{x^2} - (\overline{x})^2$$

*Эта оценка смещенная! Чтобы убедиться в этом, обозначим m = MX и воспользуемся тем, что D*x не изменяется при вычитании константы:*

$$D_x^* = D^*(x-m) = \overline{(x-m)^2} - (\overline{x-m})^2 = \overline{(x-m)^2} - (\overline{x}-m)^2$$

$$MD_x^* = M\overline{(X-m)^2} - M(\overline{X}-m)^2 = M\left(\frac{\sum (X_i-m)^2}{n}\right) - M(\overline{X}-m)^2 = \frac{\sum M(X_i-m)^2}{n} - M(\overline{X}-m)^2 =$$

$$= \frac{\sum DX}{n} - D\overline{X} = \frac{nDX}{n} - D\overline{X} = DX - D\overline{X} = DX - \frac{1}{n}DX = \frac{n-1}{n}DX \neq DX$$

Несмещённая оценка для дисперсии DX обозначается s^2 и вычисляется по формуле

$$s^2 = \frac{n}{n-1} D_X^*$$

Точечная оценка коэффициента асимметрии находится по формуле

$$A_X^* = \frac{1}{\left(D_x^*\right)^{3/2}}(x-\overline{x})^3 =$$

$$= \frac{1}{\left(D_x^*\right)^{3/2}}\left(\overline{x^3} - 3\cdot\overline{x^2}\cdot\overline{x} + 2\cdot(\overline{x})^3\right)$$

Точечная оценка коэффициента эксцесса находится по формуле

$$E_X^* = \frac{1}{\left(D_x^*\right)^2}(x-\overline{x})^4 - 3 =$$

$$= \frac{1}{\left(D_x^*\right)^2}\left(\overline{x^4} - 4\cdot\overline{x^3}\cdot\overline{x} + 6\cdot\overline{x^2}\cdot(\overline{x})^2 - 3\cdot(\overline{x})^4\right) - 3$$

Точечная оценка коэффициента корреляции случайных величин **X** и **Y** находится по формуле

$$r_{XY}^* = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sqrt{D_x^*}\cdot\sqrt{D_y^*}}$$

Задача. По двумерной таблице частот найти $\overline{x}, \overline{y}, D_x^*, D_y^*, r_{XY}^*$

X\Y	-1	0	2
0-20	5	3	4
20-40	2	1	5

$\overline{x} = \frac{10\cdot 12 + 30\cdot 8}{20} = \ldots,$

$\overline{y} = \ldots,$

$\overline{x^2} = \frac{10^2\cdot 12 + \ldots}{20} = \ldots, D_x^* = \overline{x^2} - (\overline{x})^2 = \ldots,$

$\overline{y^2} = \ldots, D_y^* = \overline{y^2} - (\overline{y})^2 = \ldots$

$\overline{xy} = \frac{10\cdot(-1)\cdot 5 + \ldots}{20} = \ldots; r_{xy}^* = \ldots$

Масштабное преобразование

Вычисление точечных оценок можно существенно упростить, если использовать масштабное преобразование.

$$Если\ \mathrm{u} = \frac{\mathrm{x} - \mathrm{a}}{\mathrm{b}},\ то$$
$$\bar{\mathrm{x}} = \bar{u} \cdot b + a,$$
$$\mathrm{D}_x^* = b^2 \cdot \mathrm{D}_u^*$$
$$\mathrm{A}_x^* = \mathrm{A}_u^*$$
$$\mathrm{E}_x^* = \mathrm{E}_u^*$$

$$Если\ \mathrm{u} = \frac{\mathrm{x} - \mathrm{a}}{\mathrm{b}},\ \mathrm{v} = \frac{\mathrm{y} - \mathrm{c}}{\mathrm{d}},\ то$$
$$\mathrm{r}_{xy}^* = \mathrm{r}_{uv}^*$$

Подбирая удобные a, b, можно перейти от чисел x к более простым числам u

Пример 1.

Для выборки из примера 1 (негруппированной)

X_i	20,1	20,3	20,4	20,5
n_i	5	5	4	6

Выберем u=(x-20,3)/0,1. Тогда

U_i	-2	0	1	2
n_i	5	5	4	6

$$\bar{u} = \ldots \Rightarrow \bar{\mathrm{x}} = \bar{u} \cdot b + a = \ldots,$$
$$\mathrm{D}_u^* = \ldots \Rightarrow \mathrm{D}_x^* = b^2 \cdot \mathrm{D}_u^* = \ldots$$
$$\mathrm{A}_u^* = \ldots \Rightarrow \mathrm{A}_x^* = \mathrm{A}_u^* = \ldots$$
$$\mathrm{E}_u^* = \ldots \Rightarrow \mathrm{E}_x^* = \mathrm{E}_u^* = \ldots$$

Метод максимального правдопообия для получения точечных оценок

1 случай. Случайная величина **непрерывного** типа.

Плотность распределения f(x,θ) зависит от параметра θ. Требуется по выборке $x_1,x_2,\ldots$ найти приближённое значение θ.

1) Составим функцию правдоподобия $L(\theta)= f(x_1,\theta)\cdot f(x_2,\theta)\cdot f(x_3,\theta)\cdot\ldots$
2) Вычислим $\ln(L(\theta))= \ln(f(x_1,\theta))+\ln(f(x_2,\theta))+\ln(f(x_3,\theta))+\ldots$
3) Найдём, в какой точке функция L(θ) (или ln(L(θ))) имеет максимум. Эта точка и будет приближённым значением для θ. Обычно для поиска максимума этого приравнивают нулю производную по θ.

Пример.

$F(x,A)=x^A$ внутри отрезка [0;1] и равна нулю вне его. Требуется найти приближённо A по выборке 0,4; 0,5; 0,3; 0,4; 0,8 методом максимального правдоподобия.

Решение. $f(x,A)= Ax^{A-1}$ внутри отрезка [0;1].

$L(A) =\ldots$

$\ln L(A) =\ldots$

$(\ln L(A))' =\ldots=0 \rightarrow A=\ldots$

2 случай. Случайная величина **дискретного** типа.

Вероятность распределения P(X=k) зависит от параметра θ. Требуется по выборке $x_1,x_2,\ldots$ найти приближённое значение θ.

1) Составим функцию правдоподобия $L(\theta) = P(X=x_1)\cdot P(X=x_2)\cdot P(X=x_3)\cdot\ldots$
2) Вычислим $\ln(L(\theta))= \ln(P(X=x_1))+\ln(P(X=x_2))+\ln(P(X=x_3))+\ldots$
3) Найдём, в какой точке функция L(θ) (или ln(L(θ))) имеет максимум. Эта точка и будет приближённым значением для θ. Обычно для поиска максимума этого приравнивают нулю производную по θ.

Пример.

X распределена по закону Пуассона. Требуется найти приближённо λ по выборке 1; 3; 3; 5; 2 методом максимального правдоподобия.

Решение.

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$
$$\mathrm{L}(\lambda) = \ldots$$
$$\ln\left(\mathrm{L}(\lambda)\right) = \ldots$$
$$0 = \ln\left(\mathrm{L}(\lambda)\right)' = \ldots \Rightarrow \lambda = \ldots$$

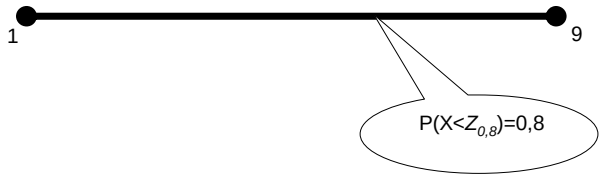
Квантиль закона распределения

Пусть X – случайная величина непрерывного типа.
Квантиль Z_a- это такое число, что

$$P(X < Z_p) = F(Z_p) = p$$

Пример.

X распределена равномерно на отрезке [1;9]. Чему равен квантиль $Z_{0,8}$?



Пример.

Квантиль стандартного нормального закона N(0;1) обозначается u_p.
Другими словами, $\Phi(u_p)=p$.
Свойство:

$$u_{1-\alpha} = -u_{\alpha}$$

p	**0.9**	**0.95**	**0.975**	**0.99**	**0.995**	**0.999**	**0.9995**
u_p	**1.282**	**1.645**	**1.960**	**2.325**	**2.576**	**3.090**	**3.291**

Вопрос.

Чему равно $u_{0,01}$?

Закон распределения "хи-квадрат"

Распределение χ^2 *с k степенями свободы* - это закон распределения случайной величины $\chi^2(k)$, равной сумме квадратов k независимых СВ U, распределенных по стандартному нормальному закону N(0,1):

$$\chi^2 (k) = U_1^2 + U_2^2 + U_3^2 + ... + U_k^2.$$

Квантиль распределения порядка p обозначается χ_p^2 (k).

Теорема 1. Пусть $x_1, x_2, ..., x_n$ – выборка из генеральной совокупности, распределенной по нормальному закону N(m,σ). Пусть $\bar{x}$ и S^2 - соответственно *выборочное среднее* и *несмещенная оценка для дисперсии*. Тогда

1. Статистики $\bar{x}$ и S^2 - независимые случайные величины;
2. Случайная величина $\frac{(n-1)S^2}{\sigma^2}$ распределена по закону $\chi^2(n-1)$;
3. Случайная величина $\frac{\bar{x} - m}{\sigma/\sqrt{n}}$ распределена по закону N(0,1).

Квантили χ_p^2 (k).

k \ p	0.005	0.01	0.05	0.1	0.9	0.95	0.99	0.995
1	0	0.0002	0.004	0.0158	2.71	3.84	6.63	7.88
4	0.2	0.297	0.711	1.06	7.78	9.49	13.3	18.5
9	1.73	2.09	3.33	4.17	14.7	16.9	21.7	23.6
19	6.84	7.63	10.1	11.7	27.2	30.1	36.2	38.6
29	13.1	14.3	17.7	19.8	39.1	42.6	49.6	52.3
40	20.7	22.2	26.5	29.1	51.8	55.8	63.7	66.8
50	28.0	29.7	34.8	37.7	63.2	67.5	76.2	79.5
75	47.2	49.5	56.1	59.8	91.1	96.2	106.4	110.3
100	67.3	70.1	77.9	82.4	118.5	124.3	135.6	140.2

Закон распределения Стъюдента

Распределение Стъюдента с k степенями свободы - это закон распределения СВ T(k), равной:

$$T(k) = \frac{U}{\sqrt{\chi^2(k)/k}}$$

где U распределена по закону N(0,1), $\chi^2(k)$ – по закону $\chi^2(k)$, U и $\chi^2(k)$ независимы.

Теорема 2. Пусть $\bar{x}$ *и* S^2 - те же, что и в предыдущей теореме. Тогда случайная величина $\frac{\bar{x} - m}{s/\sqrt{n}}$ распределена по закону T(n-1).

Квантиль распределения порядка p обозначается $t_p(k)$.

$$t_{1-\alpha}(k) = -t_\alpha(k)$$

k \ p	0.900	0.950	0.975	0.990	0.995	0.999
1	3.078	6.314	12.706	31.821	63.657	318
4	1.533	2,132	2.776	3.747	4.604	7.173
9	1.383	1.833	2.262	2.281	3.250	4.297
19	1.328	1.729	2.093	2.539	2.861	3.579
29	1.311	1.699	2.045	2.462	2.758	3.398
40	1.303	1.684	2.021	2.423	2.704	3.307
60	1.296	1.671	2.000	2.390	2.660	3.232
120	1.289	1.658	1.980	2.358	2.617	3.160
∞	1.282	1.645	1.96	2.326	2.576	3.09

$$t_p(k) \approx u_\alpha \text{ при больших } k$$

Закон распределения Фишера

Распределение Фишера с k_1 *и* k_2 *степенями свободы* – это закон распределения случайной величины $F(k_1, k_2)$, равной:

$$F(k_1, k_2) = \frac{\chi^2(k_1)/k_1}{\chi^2(k_2)/k_2}$$

где $\chi^2(k_1)$ – СВ, распределенная по закону $\chi^2(k_1)$, $\chi^2(k_2)$ – СВ, распределенная по закону $\chi^2(k_2)$ и эти две СВ независимы

Теорема 2. Пусть есть две выборки объемов n_1 и n_2 из генеральных совокупностей, распределенных по законам $N(m_i,\sigma_i)$ (i=1,2). Тогда: .

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = F(n_1 - 1, n_2 - 1)$$

Квантиль распределения порядка p обозначается $F_p(k_1,k_2)$.

$$F_{1-\alpha}(k_1, k_2) = \frac{1}{F_\alpha(k_2, k_1)}$$

Квантили распределения $F_{0,95}(k_1,k_2)$

n_1 \ n_2	1	2	3	4	5	10	15	20	30	120
1	40	8,53	5,54	4,54	4,06	3,29	3,07	2,97	2,88	2,75
2	49,5	9	5,46	4,32	3,78	2,92	2,7	2,59	2,49	2,35
3	53,6	9,16	5,39	4,19	3,62	2,73	2,49	2,38	2,28	2,13
4	55,8	9,24	5,34	4,11	3,52	2,61	2,36	2,25	2,14	1,99
10	60,2	9,39	5,23	3,92	3,3	2,32	2,06	1,94	1,82	1,65
20	61,7	9,44	5,18	3,84	3,21	2,2	1,92	1,79	1,67	1,48
30	62,2	9,46	5,17	3,82	3,17	2,16	1,87	1,74	1,61	1,41
60	62,8	9,47	5,15	3,79	3,14	2,11	1,82	1,68	1,54	1,32

Доверительные интервалы

При статистической обработке результатов наблюдений обычно требуется не только найти оценку для неизвестного параметра закона распределения изучаемой случайной величины, но и описать точность этой оценки (насколько ей можно “доверять”).

Определение. *Доверительным интервалом* (ДИ) для неизвестного параметра Θ называется интервал (Θ_1,Θ_2), содержащий (накрывающий) точное значение Θ с заданной вероятностью $\gamma = 1-\alpha$:

$$P(\Theta < \Theta < \Theta) = \gamma = 1 - \alpha.$$

Число γ *называется доверительной вероятностью,* а число α - *уровнем значимости*.

Обычно выбирают α = 0.1, 0.05 или 0.01.

Доверительные интервалы в случае нормального закона распределения генеральной совокупности

1. Доверительный интервал для m=MX при известной дисперсии σ^2.

Пусть x=(x_1, x_2,..., x_n) - выборка. Тогда

$$U = \frac{\overline{X} - m}{\sigma/\sqrt{n}} \sim N(0;1)$$

Поэтому

$$P(u_{\alpha/2} < \frac{\overline{X} - m}{\sigma/\sqrt{n}} < u_{1-\alpha/2}) = \Phi(u_{1-\alpha/2}) - \Phi(u_{\alpha/2}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

Получим отсюда неравенство для m=MX: …<m<…

$$\overline{x} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2} < m < \overline{x} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2}$$

2. Доверительный интервал для m=MX при неизвестной дисперсии σ^2.

Пусть x=(x_1, x_2,..., x_n) - выборка. Тогда

$$U = \frac{\overline{X} - m}{s/\sqrt{n}} \sim T(n-1)$$

Поэтому

$$P(t_{\alpha/2}(n-1) < \frac{\overline{X} - m}{s/\sqrt{n}} < t_{1-\alpha/2}(n-1)) = \ldots = 1 - \alpha$$

Получим отсюда неравенство для m=MX: …<m<…

$$\overline{x} - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) < m < \overline{x} + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1)$$

3. Доверительный интервал для σ^2.

Пусть x=(x_1, x_2,..., x_n) - выборка. Тогда

$$U = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

Поэтому

$$P(\chi^2_{\alpha/2}(n-1) < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{1-\alpha/2}(n-1)) = \ldots = 1 - \alpha$$

Получим отсюда неравенство для σ^2: …< σ^2 <…

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}$$

Задача.

Станок-автомат заполняет пакеты чипсами по 250 г. Считается, что станок требует подналадки, если стандартное отклонение от номинального веса превышает 5 г. Контрольное взвешивание 10 пакетов дало следующие результаты:
245,248,250,250,252,256,243,251,244,253.

1. Построить 90% доверительный интервал для стандартного отклонения от номинального веса.

2. Требует ли станок подналадки?

Решение.

Нас интересует вес пакета. Можно ли считать, что он распределён по нормальному закону?

Что называется стандартным отклонением? Какой из трёх доверительных интервалов используется в этой задаче?

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}$$

Чему равно n?

Чему равно α ?

Нужно вычислить S^2

$$u = \frac{x-250}{1};\quad \bar{u} = -0{,}8;\quad \overline{u^2} = 16{,}4;\quad D^*_u = 16{,}4 - (-0{,}8)^2 = 15{,}76$$

$$\bar{x} = 250 - 0{,}8 = 249{,}2;\quad D^*_x = 1^2 \cdot D^*_u = 15{,}76;\quad S^2 = \frac{10}{9} D^*_x = 17{,}51$$

$$\chi^2_{1-\alpha/2}(n-1) = \chi^2_{0{,}95}(9) = \ldots;\quad \chi^2_{\alpha/2}(n-1) = \chi^2_{0{,}05}(9) = \ldots$$

$$\frac{9 \cdot 17{,}51}{16{,}9} < \sigma^2 < \frac{9 \cdot 17{,}51}{3{,}33}$$

$$9{,}325 < \sigma^2 < 47{,}33$$

$$3{,}054 < \sigma < 6{,}879$$

Так нужна ли подналадка?

Доверительные интервалы в случае биномиального закона распределения генеральной совокупности

Рассмотрим схему Бернулли. Пусть случайная величина X равна количеству успехов в N испытаниях, p - вероятность успеха в одном испытании, q=1–p - вероятность неудачи. Пусть x=(x_1, x_2,..., x_n) – выборка. Тогда общее число испытаний равно N·n,

$\nu = \frac{\sum x_i}{Nn} = \frac{\bar{x}}{N}$ *– частота успехов.*

По *интегральной теореме Муавра-Лапласа* при *большом числе испытаний* n приближённо

$$\nu \sim N(p, \sqrt{pq/Nn})$$

Отсюда получается приближённый доверительный интервал для p:

$$\nu - \frac{\sqrt{\nu(1-\nu)}}{\sqrt{nN}} \cdot u_{1-\alpha/2} < p < \nu + \frac{\sqrt{\nu(1-\nu)}}{\sqrt{nN}} \cdot u_{1-\alpha/2}$$

Задача.

С автоматической линии, производящей подшипники, было отобрано 400 штук, причём 10 оказались бракованными.

Найти 95%-й доверительный интервал для доли брака в производимой линией продукции

Решение.

Нас интересует количество бракованных подшипников. Можно ли считать, что оно распределено по биномиальному закону?

Чему равно N?

Чему равно n?

Чему равно α ?

Чему равна частота успехов ν?

$$\nu - \frac{\sqrt{\nu(1-\nu)}}{\sqrt{nN}} \cdot u_{1-\alpha/2} < p < \nu + \frac{\sqrt{\nu(1-\nu)}}{\sqrt{nN}} \cdot u_{1-\alpha/2}$$

Доверительный интервал для MX в случае произвольного закона распределения и большого объема выборки

Пусть x=(x_1, x_2,..., x_n) - выборка из *произвольно распределённой генеральной совокупности* и *n велико*. Пусть DX=σ^2 . По *центральной предельной теореме* приближённо

$$\frac{(\overline{X}-m)}{(\sigma/\sqrt{n})} \sim N(0,1)$$

Поэтому ***приближенный*** ДИ для MX получается таким же, как и в случае нормального распределения генеральной совокупности (см. выше).

$$\overline{x}-\frac{\sqrt{DX}}{\sqrt{n}}\cdot u_{1-\alpha/2} < m < \overline{x}+\frac{\sqrt{DX}}{\sqrt{n}}\cdot u_{1-\alpha/2}$$

Доверительный интервал для параметра в случае генеральной совокупности, распределенной по закону Пуассона

Пусть x=(x_1, x_2,..., x_n) - выборка из *генеральной совокупности, распределённой по закону Пуассона* и *n велико*. Тогда MX=DX=λ. Применяя формулу из предыдущего пункта, получаем:

$$\overline{x}-\frac{\sqrt{\overline{x}}}{\sqrt{n}}\cdot u_{1-\alpha/2} < \lambda < \overline{x}+\frac{\sqrt{\overline{x}}}{\sqrt{n}}\cdot u_{1-\alpha/2}$$

Проверка статистических гипотез

Статистической гипотезой (СГ) называется предположение **о параметрах** или **законе распределения** случайной величины. Проверяемая гипотеза называется *нулевой* и обозначается H_0. Наряду с ней рассматривается одна из *конкурирующих гипотез* H_1, H_2 и т.п.

Пример. Пусть $\mathbf{H_0}$ имеет вид: “неизвестный параметр Θ равен числу Θ_0 ”, т.е.

$$\mathbf{H_0: \Theta = \Theta_0}\ .$$

В этом случае возможные конкурирующие гипотезы:

$$\mathbf{H_1: \Theta \neq \Theta_0;}$$
$$\mathbf{H_1: \Theta < \Theta_0;}$$
$$\mathbf{H_1: \Theta > \Theta_0.}$$

Правило, по которому принимается решение о принятии гипотезы, называется *критерием*. Решение принимается на основании наблюдения подходящей случайной величины Z, называемой *статистикой критерия*.

Порядок проверки статистической гипотезы:

1. По критерию строится *область принятия гипотезы* V;
2. По выборке находится *выборочное значение* $\mathbf{Z_в}$ случайной величины Z;
3. Если это значение попало в область V , то гипотеза $\mathbf{H_0}$ принимается, иначе отвергается (*принимается конкурирующая*).

Принятие решения в условиях неопределенности естественно может привести к ошибке.

Ошибка первого рода: *гипотеза истинна, а мы ее отвергли*.
Вероятность этого обозначается α и называется *уровень значимости*:

$$\mathbf{P(Z \notin V \mid H_0) = \alpha.}$$

Ошибка второго рода: *гипотеза ложна, а мы ее приняли*.
Вероятность этого обозначается **β**; число **1–β** называется *мощность критерия*:

$$\mathbf{P(Z \in V \mid H_1)=\beta.}$$

Проверка гипотез о параметрах закона распределения

Для проверки такой гипотезы следует:
1.Сформулировать исходную гипотезу в виде H_0: $\Theta = \Theta_0$.
2.Выбрать конкурирующую гипотезу из списка, приведенного в примере.
3.Выбрать статистику Z критерия.
4.В зависимости от вида конкурирующей гипотезы построить область V по одной из формул
(здесь буквой k обозначены соответствующие квантили).

$$H_1 : \Theta \neq \Theta_0 \Rightarrow V = \{z : k_{\alpha/2} < z < k_{1-\alpha/2}\}$$

$$H_1 : \Theta < \Theta_0 \Rightarrow V = \{z : z > k_{\alpha}\}$$

$$H_1 : \Theta > \Theta_0 \Rightarrow V = \{z : z < k_{1-\alpha}\}$$

5. По выборке вычислить $\mathbf{Z_в}$.
6. Если это значение попало в область V , то гипотеза H_0 принимается, иначе отвергается (принимается конкурирующая).

Таким образом, чтобы описать критерий проверки параметрической гипотезы, достаточно указать формулу для подсчета $\mathbf{Z_в}$ и какие квантили следует использовать.

Замечание. Для любой из написанных выше формул для V легко проверить справедливость соотношения $\mathbf{P(Z \notin V | H_0) = \alpha}$.

Проверка гипотез о параметрах нормального закона распределения

Пусть $\mathbf{x=(x_1, x_2,..., x_n)}$ - выборка.

1. H_0 : $m = m_0$ и дисперсия σ^2 известна.
В этом случае

$$z^{‰} = \frac{(\bar{x} - m_0)}{(\sigma/\sqrt{n})}; k_p = u_p.$$

2. H_0 : $m = m_0$ и дисперсия σ^2 неизвестна.
В этом случае

$$z^{в} = \frac{(\bar{X} - m_0)}{(s/\sqrt{n})}; k_p = t_p(n-1).$$

3. H_0: $\sigma = \sigma_0$ и математическое ожидание неизвестно.
В этом случае

$$z^{в} = \frac{(n-1)s^2}{\sigma_0^2}; k_p = \chi_p^2(n-1)$$

Пусть теперь две случайные величины, X_1 и X_2 и соответственно две выборки из соответствующих генеральных совокупностей, распределённых по законам $\mathbf{N(m_1, \sigma_1)}$ и $\mathbf{N(m_2, \sigma_2)}$.

4. H_0 : $m_1 = m_2$ и дисперсии известны. В этом случае

$$z^{в} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; k_p = u_p.$$

5. H_0 : $\sigma_1 = \sigma_2$ и математические ожидания неизвестны.
Тогда

$$z^{в} = s_1^2 / s_2^2; \quad k_p = F_p(n_1 - 1, n_2 - 1).$$

Проверка гипотез о параметре p биномиального закона распределения

Рассмотрим схему Бернулли. Другими словами,
X - количество успехов в N испытаниях, p – вероятность успеха в одном испытании, q=1-p – вероятность неудачи.
Пусть величина X измеряется n раз (так что всего испытаний N·n). Тогда частота успехов равна

$$\nu = \sum x_i \Big/ nN = \bar{x} \Big/ N$$

1. $H_0 : p=p_0$. В этом случае

$$Z^{в} = \frac{\nu - p_0}{\sqrt{\frac{p_0(1-p_0)}{Nn}}}; k_p = u_p.$$

Пусть теперь две серии испытаний и соответственно две частоты успехов ν_1 и ν_2. Обозначим через ν **общую частоту успехов:**

$$\nu = \frac{\text{всего успехов}}{\text{всего испытаний}} = \frac{N_1 n_1 \nu_1 + N_2 n_2 \nu_2}{N_1 n_1 + N_2 n_2}$$

2. $H_0 : p_1=p_2$. В этом случае

$$z^{в} = \frac{\nu_1 - \nu_2}{\sqrt{\nu(1-\nu)\left(\frac{1}{N_1 n_1} + \frac{1}{N_2 n_2}\right)}}; k_p = u_p.$$

Задача.

Имеются следующие результаты испытания новой прививки:

Делали прививку		Не делали прививку	
Заболели	Нет	Заболели	Нет
3	147	12	188

Доказывают ли эти данные эффективность прививки? α=0,01

Решение.

X_1, X_2 – количества заболевших. Биномиальный закон.

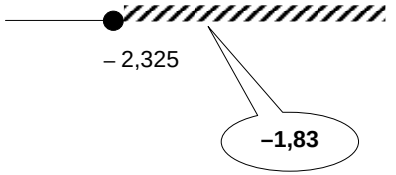
p_1 – вероятность заболеть, сделав прививку,
p_2 – вероятность заболеть, не сделав прививку.

H_0: $p_1 = p_2$ (т.е. прививка неэффективна),
H_1: $p_1 < p_2$ (т.е. прививка эффективна)

$$\nu_1 = \frac{3}{150} = 0{,}02,\ \nu_2 = \frac{12}{200} = 0{,}06 \qquad \nu = \frac{3+12}{150+200} \approx 0{,}043$$

$$z^{в} = \frac{\nu - p_0}{\sqrt{\nu(1-\nu)\left(\frac{1}{N_1 n_1} + \frac{1}{N_2 n_2}\right)}}; k_p = u_p.$$

$$Z_{в} = \frac{0{,}02 - 0{,}06}{\sqrt{0{,}043 \cdot (1-0{,}043)\left(\frac{1}{150} + \frac{1}{200}\right)}} \approx -1{,}83$$



H_1: $p_1 < p_2 \Rightarrow V=\{Z > k_\alpha\} = \{Z > u_{0,01}\} = \{Z > -2{,}325\}$

Ответ. H_0 принимается, т.е. **недостаточно оснований считать, что прививка эффективна**.

Проверка гипотез о значимости коэффициента корреляции

$H_0 : r_{XY}=0$

$$z^{в} = r^{*}_{XY}; k_p = \frac{t_p(n-2)}{\sqrt{n-2+\left(t_p(n-2)\right)^2}}$$

Пример. *Предполагается наличие приближённо линейной зависимости между доходами госбюджета и расходами на социальное обеспечение некоторой стране за 10 лет. По статистическим данным рассчитан $r^*_{XY} = 0,42$. Проверить гипотезу на уровне $\alpha=0,01$*

Проверка гипотез о виде функции распределения (критерий хи-квадрат)

Пусть имеется выборка **x=(x_1, x_2,..., x_n).** Проверяется гипотеза о том, что случайная величина X распределена по какому-то конкретному закону.

Опишем порядок проверки такой гипотезы:

1. По выборке оценить значения неизвестных параметров гипотетического закона распределения (если такие есть). Так, например

Нормальный закон $N(m,\sigma)$:	$m \approx \bar{x}, \sigma \approx s;$
Показательный закон $E(\lambda)$:	$\lambda \approx 1/\bar{x};$
Закон Пуассона $P(\lambda)$:	$\lambda \approx \bar{x};$
Биномиальный закон $B(N,p)$:	$p \approx \bar{x}/N;$
Геометрический закон $G(p)$:	$p \approx 1/\bar{x}$

2. Всю область значений случайной величины X разбить на несколько подмножеств (интервалов) $\Delta_1, \Delta_2, \ldots, \Delta_r$.
3. Заполнить группированную таблицу частот; вторая строка таблицы содержит *наблюдаемые частоты* попадания элементов выборки в соответствующие интервалы.

Интервал	Δ_1	Δ_2	…	Δ_r
Наблюдаемые частоты	n_1	n_2	…	n_r
Ожидаемые частоты	$\tilde{n}_1$	$\tilde{n}_2$	…	$\tilde{n}_r$

4. По формулам из предполагаемого закона распределения найти вероятности $p_i=P(X\in\Delta_1)$ (i=1,...,r) попадания в интервалы. Добавить в таблицу частот еще одну строку "***ожидаемые частоты***" и заполнить ее по формулам

$$\tilde{n}_i = n \cdot p_i, \quad p_i = P(x \in \Delta_i).$$

Суммы чисел в последних двух строчках таблицы должны быть равны объему выборки!

Если некоторые ожидаемые частоты оказались меньше, чем 3, то следует объединить соседние интервалы (и столбцы таблицы)

5. Вычислить $$z^{в} = \sum_{i=1}^{r} \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i}$$

Если это число окажется меньше, чем $\chi^2_{1-\alpha}(r-L-1)$, то гипотеза о законе распределения принимается, иначе отвергается.
Здесь через r обозначено количество интервалов ,
а через L – количество неизвестных параметров закона распределения, найденных на первом шаге.

Задача.

Имеются следующие данные о количестве отказов аппаратуры за 10000 часов работы. Проверить гипотезу о том, что число отказов подчиняется закону Пуассона. Взять α=0,01

Количество отказов **X**	**0**	**1**	**2**	**3**	**4**	**5**
Частота n_i	427	235	72	21	1	1

Решение.

Закон Пуассона имеет вид $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},\ k=0,1,2,\ldots$

1. Параметр λ неизвестен, поэтому надо найти его приближённое значение $\lambda \approx \bar{x} = 451/757 \approx 0{,}6$

2,3. Область значений X в законе Пуассона представляет собой 0,1,2,…., поэтому в таблицу надо добавить столбец "≥6"

X	**0**	**1**	**2**	**3**	**4**	**5**	**≥6**
Наблюдаемая частота n_i	427	235	72	21	1	1	0
Ожидаемая частота $\tilde{n}_i$							

4. Теперь надо рассчитать ожидаемые частоты

$$\tilde{p}_0 = P(X=0)=\frac{0{,}6^0}{0!}e^{-0{,}6}=e^{-0{,}6}\approx 0{,}549;\quad \tilde{n}_0 = n\cdot\tilde{p}_0 = 757\cdot 0{,}549 \approx 416$$

$$\tilde{p}_1 = P(X=1)=\frac{0{,}6^1}{1!}e^{-0{,}6}\approx 0{,}329;\quad \tilde{n}_1 = n\cdot\tilde{p}_1 = 757\cdot 0{,}329 \approx 249;$$

$$\tilde{p}_2 = P(X=2)=\frac{0{,}6^2}{2!}e^{-0{,}6}\approx 0{,}099;\quad \tilde{n}_2 = n\cdot\tilde{p}_2 = 757\cdot 0{,}099 \approx 75...$$

X	**0**	**1**	**2**	**3**	**4**	**5**	**≥6**
Наблюдаемая частота n_i	427	235	72	21	1	1	0
Ожидаемая частота $\tilde{n}_i$	416	249	75	15	2	0	0

Некоторые ожидаемые частоты оказались меньше 3, поэтому объединяем интервалы!

X	**0**	**1**	**2**	**≥3**
n_i	427	235	72	23
$\tilde{n}_i$	416	249	75	17

5. Рассчитаем $Z^в$

$$z^в = \sum_{i=1}^{r}\frac{(n_i-\tilde{n}_i)^2}{\tilde{n}_i} = \frac{(427-416)^2}{416}+\ldots = 3{,}316$$

$$\chi^2_{1-\alpha}(r-L-1)=\chi^2_{0{,}99}(4-1-1)=9{,}21$$

Ответ: $Z^в<9{,}21$=>гипотеза о распределении по закону Пуассона принимается на уровне значимости α=0,01.

Задача.

Для следующей выборки проверить гипотеза о нормальном законе распределения. Взять α=0,01.

Δ_i	**2 — 4**	**4 — 6**	**6 — 8**	**8 — 10**	**10 — 12**
n_i	2	8	14	10	6

Решение.

Нормальный закон имеет вид $P(a<X<b)=\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right)$

1. Параметры m,σ неизвестны, поэтому надо найти их приближённые значения $m \approx \bar{x} = 7{,}5;$ $\sigma \approx S = \sqrt{S^2} \approx 2{,}2$

2,3. Область значений X в нормальном законе представляет собой всю числовую ось, поэтому в таблицу надо добавить столбцы "≤2" и "≥12"

Δ_i	≤ 2	2 — 4	4 — 6	6 — 8	8 — 10	10 — 12	≥ 12
n_i	0	2	8	14	10	6	0
$\tilde{n}_i$							

4. Теперь надо рассчитать ожидаемые частоты

$$p_1 = P(-\infty < X < 2) = \Phi\left(\frac{2-7{,}5}{2{,}2}\right) - \Phi(-\infty) = \Phi(-2{,}5) - 0 =$$

$$= 1 - \Phi(2{,}5) = 1 - 0{,}9938 = 0{,}0062;$$

$$\tilde{n}_1 = n \cdot p_1 = 40 \cdot 0{,}0062 \approx 0{,}248$$

$$p_2 = P(2 < X < 4) = \Phi\left(\frac{4-7{,}5}{2{,}2}\right) - \Phi\left(\frac{2-7{,}5}{2{,}2}\right) = \Phi(-1{,}59) - \Phi(-2{,}5) =$$

$$= 1 - \Phi(1{,}59) - (1 - \Phi(2{,}5)) =$$

$$= 0{,}0497; \quad n_2 = n \cdot p_2 = 40 \cdot 0{,}0497 \approx 1{,}988...$$

Δ_i	≤ 2	2 — 4	4 — 6	6 — 8	8 — 10	10 — 12	≥ 12
n_i	0	2	8	14	10	6	0
$\tilde{n}_i$	0,248	1,988	7,696	12,07	13,2	4	0,8

Некоторые ожидаемые частоты оказались меньше 3, поэтому объединяем интервалы!

Δ_i	≤ 6	6 — 8	8 — 10	≥ 10
n_i	10	14	10	6
$\tilde{n}_i$	9,932	12,07	13,2	4,8

5. Рассчитаем $Z^в$

$$z^в = \sum_{i=1}^{r} \frac{(n_i - \tilde{n}_i)^2}{\tilde{n}_i} = \frac{(10 - 9{,}932)^2}{9{,}932} + ... = 1{,}53$$

$\chi^2_{1-\alpha}(r-L-1) = \chi^2_{0{,}99}(4-2-1) = 6{,}63$

Ответ: $Z^в < 6{,}63$ => гипотеза о нормальном законе распределения принимается на уровне значимости α=0,01.

Проверка гипотез о независимости (критерий хи-квадрат)

Пусть имеется двумерная выборка. Проверяется гипотеза о независимости двух случайных величин, **X** и **Y**.

Обозначим через $\mathbf{n}_{ij}$ количество элементов выборки, для которых $X=x_i$ и $Y=y_j$. Введем также обозначения

$$n_{\bullet j} = \sum_i n_{ij}, \quad n_{i\bullet} = \sum_j n_{ij}$$

Проверяется гипотеза о независимости СВ X и Y. Если гипотеза верна, то по определению $P[X=x_i; Y=y_j] = P[X=x_i] \cdot P[Y=y_j]$.

Учитывая приближенные равенства $P[X=x_i; Y=y_j] \approx n_{ij}/n$; $P[X=x_i] \approx n_{i\bullet}/n$; $P[Y=y_j] \approx n_{\bullet j}/n$,

найдем ожидаемые частоты по формуле

$$\tilde{n}_{ij} = n \cdot P[X = x_i; Y = y_j] \approx \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

$$z^{в} = \sum_{i,j} \frac{(n_{ij} - \widetilde{n}_{ij})^2}{\widetilde{n}_{ij}} = n\left(\sum_{i,j} \frac{n_{ij}^2}{n_{i\bullet} \cdot n_{\bullet j}} - 1 \right)$$

Если это число окажется меньше, чем $\chi^2_{1-\alpha}((n_X-1)(n_Y-1))$, то гипотеза о независимости X и Y принимается, иначе отвергается.
Здесь n_X и n_Y - соответственно количество строк и столбцов в двумерной таблице частот.

Задача. *Имеются следующие данные:*

Работа за компьютером	Динамика состояния зрения	
	Не ухудшилось	Ухудшилось
Не работает	70	5
Недавно работает	60	20
Давно работает	10	45

Есть ли зависимость? Взять α=0,01

Решение.

$$Z^{в} = 210 \cdot \left(\frac{70^2}{75 \cdot 140} + \ldots - 1 \right) =$$

$$= 210 \cdot \left(\frac{70^2}{75 \cdot 140} + \frac{5^2}{75 \cdot 70} + \frac{60^2}{80 \cdot 140} \ldots - 1 \right) = 84{,}68$$

$$\chi^2_{1-\alpha}\left((n_x - 1) \cdot (n_y - 1)\right) = \chi^2_{0{,}99}(2 \cdot 1) = 9{,}21$$

Ответ. *Есть зависимость.*

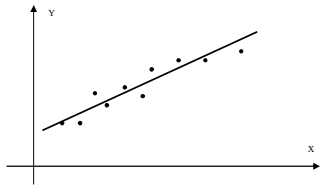
Метод наименьших квадратов

Во многих случаях возникает задача: *найти кривую заданного вида, наиболее точно приближающую экспериментальные данные*.
Математически это формулируется так:
требуется подобрать такие значения параметров β, чтобы график функции

$$f(x) = \beta_1\varphi_1(x) + \beta_2\varphi_2(x) + \ldots + \beta_m\varphi_m(x)$$

проходил как можно ближе к заданным точкам (x_i,y_i), i=1,2..n. *Здесь* β_i *– неизвестные коэффициенты,* φ_i *– известные функции.*

Пример (на рис).
Требуется подобрать прямую $y = \beta_1 + \beta_2 x$, наиболее близкую к заданным точкам.



Чтобы такая задача могла считаться корректно сформулированной, нужно как-то конкретизировать понятие *близости* точек к кривой. Это можно сделать многими способами, наиболее распространён следующий:

Для заданных точек (x_i,y_i), i=1,2..n,
найти такие значения параметров β_i,
чтобы минимизировать остаточную сумму квадратов:

$$Q(\beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \to \min \qquad (*)$$

Оказывается, что в этом случае получаются удобные формулы для подсчёта коэффициентов *β*.
Дифференцируя *Q(β)*, и приравнивая производную нулю, легко получить следующую систему линейных уравнений для нахождения коэффициентов β_i:

$$\begin{cases} A_{11}\beta_1 + & A_{12}\beta_2 + \ldots & A_{1m}\beta_m & = B_1 \\ A_{21}\beta_1 + & A_{22}\beta_2 + \ldots & A_{2m}\beta_m & = B_2 \\ \ldots & \ldots & \ldots & \ldots \\ A_{m1}\beta_1 + & A_{m2}\beta_2 + \ldots & A_{mm}\beta_m & = B_m \end{cases}$$

где

$$A_{kl} = \sum_{i=1}^{n} \varphi_k(x_i)\varphi_l(x_i), \quad B_k = \sum_{i=1}^{n} y_i \varphi_k(x_i)$$

Пример. Имеются следующие экспериментальные данные

№ п/п	1	2	3	4	5	6	7	8
X	2	3	2	2	4	3	5	5
Y	1	2	4	3	4	5	7	9

Для этих данных требуется подобрать наилучшую параболу
$y = \beta_1 + \beta_2 x + \beta_3 x^2$ методом наименьших квадратов

Здесь $\varphi_1(x)=1$, $\varphi_2(x)=x$, $\varphi_3(x)=x^2$. *Вычисляем коэффициенты системы уравнений:*

$$A_{11} = \sum \varphi_1\varphi_1 = \sum 1 = 8, \quad A_{12} = A_{21} = \sum \varphi_1\varphi_2 = \sum x = 26,$$
$$A_{13} = A_{31} = \sum \varphi_1\varphi_3 = \sum x^2 = 96, \quad A_{22} = \sum \varphi_2\varphi_2 = \sum x^2 = 96,$$
$$A_{23} = A_{32} = \sum \varphi_2\varphi_3 = \ldots, \quad A_{33} = \ldots.$$

$$B_1 = \sum y\varphi_1 = \sum y = 35, \quad B_2 = \sum (y\varphi_2) = \sum (yx) = \ldots, B_3 = \ldots$$

Получаем систему уравнений

$$\begin{cases} 8\,\beta_1 + 26\,\beta_2 + 96\,\beta_3 = 35, \\ 26\,\beta_1 + 96\,\beta_2 + 392\,\beta_3 = 133, \\ 96\,\beta_1 + 392\,\beta_2 + 1716\,\beta_3 = 559, \end{cases}$$

откуда $\beta_1 = 6{,}15$, $\beta_2 = -3{,}06$, $\beta_3 = 0{,}68$.

Таким образом, наилучшее МНК-приближение в виде параболы для исходных данных имеет вид $y = 6{,}15 - 3{,}06x + 0{,}68x^2$.

Интересно сравнить экспериментальные и расчётные значения для этой формулы:

x_i	2	3	2	2	4	3	5	5
y_i	1	2	4	3	4	5	7	9
$y_{i,расч}$	2,76	3,1	2,76	2,76	4,8	3,1	7,87	7,87

Линейная регресия

В частном случае, когда по МНК ищется прямая линия $y=\beta_1+\beta_2 x$, получаются легко запоминающиеся формулы:

$$A_{11} = \sum \varphi_1\varphi_1 = \sum 1 = n, \; A_{12} = A_{21} = \sum \varphi_1\varphi_2 = \sum x, \; A_{22} = \sum \varphi_2\varphi_2 = \sum x^2$$
$$B_1 = \sum y\varphi_1 = \sum y, \; B_2 = \sum (y\varphi_2) = \sum (yx)$$

$$\begin{cases} n\beta_1 + \left(\sum x\right)\beta_2 = \sum y \\ \left(\sum x\right)\beta_1 + \left(\sum x^2\right)\beta_2 = \sum xy \end{cases} \Rightarrow \begin{cases} \beta_1 + \bar{x}\beta_2 = \bar{y} \\ \bar{x}\beta_1 + \overline{x^2}\beta_2 = \overline{xy} \end{cases}$$

Линейная регрессия “y на x”

$$y = \beta_1 + \beta_2 x \Rightarrow \beta_2 = \frac{\begin{vmatrix} 1 & \bar{y} \\ \bar{x} & \overline{xy} \end{vmatrix}}{\begin{vmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{vmatrix}} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2}, \quad \beta_1 = \bar{y} - \beta_2\bar{x}$$

Наряду с регрессией y на x рассматривают также регрессию x на y:

$$x = \tilde{\beta}_1 + \tilde{\beta}_2 y \Rightarrow \tilde{\beta}_2 = \frac{\begin{vmatrix} 1 & \overline{y} \\ \overline{x} & \overline{xy} \end{vmatrix}}{\begin{vmatrix} 1 & \overline{y} \\ \overline{y} & \overline{y^2} \end{vmatrix}} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{y^2} - \left(\overline{y}\right)^2}, \quad \tilde{\beta}_1 = \overline{x} - \tilde{\beta}_2 \overline{y}$$

Линейная регрессия "x на y"

Эти две прямые всегда пересекаются в точке $(\overline{x}, \overline{y})$. Совпадают они только если все точки (x_i,y_i) лежат на одной прямой.

Качество линейного приближения оценивают, вычислив

$$ESS = \sum \left(y_i - y_{i,расч}\right)^2$$

$$RSS = \sum \left(y_{i,расч} - \overline{y}\right)^2$$

$$F = \frac{(n-2) \cdot RSS}{ESS}$$

Если $F > F_{1-\alpha}(1, n-2)$, то линейная регрессия даёт хорошее приближение ("является значимой")

Задача. Имеются следующие экспериментальные данные за 7 лет о среднем доходе (x) и среднем потреблении (y):

Годы	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Доход	14,5	15,7	16,3	18,5	20,3	21,7	23
Потребление	12	12,7	13	15,5	16,5	17,3	20

Для этих данных требуется выписать уравнения линейной регрессии

Решение

$$\overline{x} \approx 18{,}57, \quad \overline{y} \approx 15{,}29, \quad \overline{xy} \approx 240{,}87,$$

$$\overline{x^2} \approx 353{,}81, \quad \overline{y^2} \approx 240{,}87$$

$$\beta_2 \approx \frac{240{,}87 - 18{,}57 \cdot 15{,}29}{353{,}81 - 18{,}57^2} = 0{,}89, \quad \beta_1 \approx 15{,}29 - 0{,}89 \cdot 18{,}57 = -1{,}17$$

$$\tilde{\beta}_2 \approx \frac{240{,}87 - 18{,}57 \cdot 15{,}29}{240{,}87 - 15{,}29^2} = 1{,}09, \quad \tilde{\beta}_1 \approx 18{,}57 - 1{,}09 \cdot 15{,}29 = 1{,}85$$

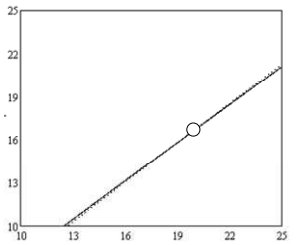


Диаграмма рассеивания

Таким образом, уравнения линейной регрессии имеют вид
y = –1,17+0,89 x и x = 1,85+1,09 y
На потребление тратится примерно 89% дохода.

Проверим теперь значимость полученного уравнения линейной регрессии y на x.

x_i	14,5	15,7	16,3	18,5	20,3	21,7	23
y_i	12	12,7	13	15,5	16,5	17,3	20
$y_{i,расч}$	11,68	12,74	13,27	15,22	16,82	18,06	19,21

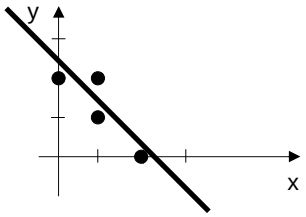
$$ESS = (12-11{,}68)^2 + (12{,}7-12{,}74)^2 + \ldots = 1{,}56$$

$$\overline{y} = 15{,}29; \quad RSS = (11{,}68-15{,}29)^2 + (12{,}74-15{,}29)^2 + \ldots = 48{,}96$$

$$F = \frac{5 \cdot 48{,}96}{1{,}56} = 157{,}3$$

Возьмём α=0,1.
$F > F_{1-\alpha}(1,n-2) = F_{0,9}(1,5)=4{,}06$ =>
построенная линейная регрессия даёт хорошее приближение зависимости y от x, является значимой на уровне значимости α=0,1.

Задача. Построить уравнение линейной регрессии y на x и проверить его значимость на уровне α=0,05



Решайте!

Ответ. 1) y = 2,25 – x

2) RSS=2, ESS=0,75,
$F=5{,}333 < F_{0,9}(1,2)=8{,}53$ => уравнение не значимо.

Список литературы

1. Колемаев В.А. и др. Теория вероятностей и математическая статистика. – М.: «Высш. школа», 1991.

2. Лагутин М.Б. Наглядная математическая статистика. – М.: .: БИНОМ. Лаборатория знаний, 2007.

3. Краснов М.Л., Киселёв А.И., Макаренко Г.И., Шикин Е.В., Заляпин В.И. Вся высшая математика – М.: Эдиториал УРСС, 2000-2002.

4. Кобзарь А.И. Прикладная математическая статистика. – М.: ФИЗМАТГИЗ, 2006.

5. Болгов В.А., Демидович Б.П., Ефимов А.В. и др. Сборник задач по математике для втузов. В 4-х частях. – М.: .: «Наука», 1993.

Оглавление

Св. план 2008 г.; поз.

Милевский Александр Станиславович

ВЫСШАЯ МАТЕМАТИКА. Ч.4. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Конспект лекций

Подписано в печать	Формат 60x84 / 16
Заказ №	Усл. печ. л. –
Тираж –	

127994, Москва, ул. Образцова, 15
Типография МИИТа