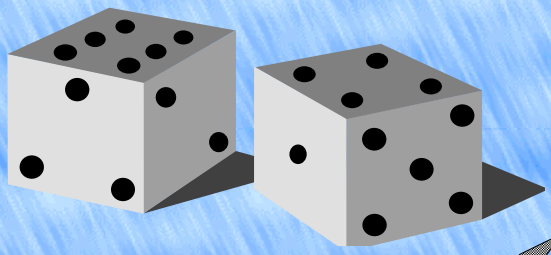
В.Н. Бандура, В.Д. Породников

Теория



вероятностей и математическая статистика

Учебное пособие для экономистов и статистиков

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Настоящее учебное пособие предназначено для студентов экономических и учетно-финансовых факультетов высших учебных заведений по курсу "Теория вероятностей и математическая статистика". Поэтому примеры и упражнения взяты из социально-экономической сферы.

В наше время, когда происходит бурный процесс математизации наших знаний, нельзя обойтись без точных количественных методов описания самых разнообразных процессов. Современная организация производства и торговли, банковского дела и экономики, биология и медицина и т.д. требуют точности и ясности изложения. Научное изложение должно быть кратким и вполне определенным. Без этого требования не может быть науки как системы знаний. Математическая символика позволяет автоматизировать проведение тех действий, которые необходимы для получения выводов, сжимать запись информации, делать ее обозримой и удобной для дальнейшей обработки.

Сейчас, в нашей высшей школе, в перечень дисциплин экономических специальностей все больше и больше включается математическая экономика. Переход к рыночной экономике обусловил необходимость подготовки специалистов, владеющих аппаратом анализа и выбора экономических вариантов на основе математико-статистических исследований, которые реализуются при помощи программного обеспечения на ЭВМ.

Пособие состоит из двух частей. Первая, которая включает основы теории вероятностей, раскрывает аксиоматическое построение теории вероятностей, на основе которого ведется изложение теории случайных событий и их вероятностей; случайные величины, их распределений и числовых характеристики; предельные теоремы теории вероятностей. Во второй части рассмотрены элементы математической статистики и некоторые ее применениям в экономических исследованиях.

Авторы благодарены рецензентам за сделанные замечания, которые в значительной мере способствовали улучшению структуры и способа изложения.

Авторы будут признательны всем заинтересованным специалистам за возможные замечания и предложения по улучшению пособия.

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

§17. Предмет математической статистики

Математическая статистика - раздел математики, посвященный методам систематизации, обработки и использования данных, изучения закономерностей массовых явлений и их взаимосвязей. Задача математической статистики состоит в том, чтобы по результатам ограниченного числа наблюдений за массовым явлением составить представление о законе его осуществления с целью последующего прогнозирования. Теория вероятностей является логической основой математической статистики, она дает возможность осмысления и интерпретации выводов, полученных исходя из экспериментальных данных.

Например, выбрав наудачу изо всей партии деталей, изготовленных неким предприятием, лишь некоторую их совокупность и проверив их качество по одному или нескольким признакам, можно сделать выводы о качестве всей партии в целом, указав при этом пределы возможных отклонений от характеристик, полученных на основе данных эксперимента (выборочных данных).

Следующие задачи математической статистики являются основными:

а) описание в той или иной целесообразной форме данных, полученных в результате эксперимента;

б) описание закона распределения исследуемого признака или признаков во всей совокупности изучаемых явлений (генеральной совокупности), в частности, оценивание ее числовых характеристик (средних, дисперсий и т. д.);

в) описание взаимосвязей между различными признаками различных массовых явлений, изучаемых совместно: установление формы связей, их степень и пр.;

г) проверка гипотез о виде искомого распределения или, если такой вид предполагается известным из предыдущего опыта, о значениях параметров, его определяющих.

Методы математической статистики используются при решении задач планирования и организации промышленного производства, при анализе демографических, социальных и экономических процессов, при контроле качества выпускаемой продукции, исследовании надежности функционирования сложных технических систем и систем управления.

При этом математическая статистика не дает рекомендаций касательно интерпретации числовых результатов обследования выборочных данных. Последнее есть задача и прерогатива тех конкретных отраслей знания, в которых используются эти результаты.

§18. Некоторые понятия и процедуры описательной статистики

Исходными понятиями математической статистики являются понятия *генеральной совокупности* и *выборки*. Под генеральной совокупностью понимают множество всех реально существующих или даже только мыслимых однородных объектов, изучаемых с некоторой общей точки зрения. Например, изучая длительность телефонного разговора, под генеральной совокупностью следует понимать множество всех абонентов телефонной сети данного района, города, области в зависимости от решаемой задачи. Это понятие, следовательно, необходимо уточнять в каждой конкретной задаче. Мы будем считать в дальнейшем, что каждый элемент генеральной совокупности описывается одной или несколькими числовыми характеристиками. С теоретико-вероятностной точки зрения генеральной совокупности ставится в соответствие случайная величина или случайный вектор.

Изучение всей генеральной совокупности возможно лишь при небольшом ее объеме и умеренных затратах, которых оно требует. Вместо полного изучения на практике прибегают к изучению только ее части, называемой *выборкой*.

Так, в указанном выше примере реально возможно фиксировать длительности телефонных разговоров некоторого конечного числа (n) клиентов, получая при этом набор чисел $t_1, t_{2,} \ldots t_n$ - длительностей разговоров, в самом деле наблюдавшихся.

В силу упомянутого выше соответствия выборку реальных (физических) объектов будем отождествлять с набором $x_1, x_{2,} \ldots x_n$ скаляров или векторов, каждый член которого есть результат наблюдения над характеристиками первого, второго ... объектов выборки.

Как уже говорилось, генеральной совокупности ставится в соответствие случайная величина или случайный вектор ξ. Поэтому выборке следует ставить в соответствие набор значений, принятых случайной величиной (вектором) ξ после n -кратного наблюдения: $x_1, x_{2,} \ldots x_n$. Имея в виду то обстоятельство, что при теоретическом изучении статистических процедур выборочные значения следует считать неизвестными и непредсказуемыми, выборку до ее осуществления на практике понимают как конечную последовательность случайных величин или векторов ($\xi_1, \xi_{2,} \ldots \xi_n$), таких, что:

- а) $\xi_1, \xi_{2,} \ldots \xi_n$ - независимы в совокупности;
- б) ξ_i (i-й член выборки) имеет тот же закон распределения, что и генеральная совокупность ξ , i =1,2,...,n.

Указанные предположения а), б) касательно выборки диктуются существом дела: результат наблюдения над любым членом выборки не должен влиять на прочие результаты, если мы хотим изучать генеральную совокупность в «чистом виде», и каждый член выборки должен нести в себе ту же информацию, что и генеральная совокупность в целом.

Итак, следует различать два случая употребления понятия «выборка»:

- выборка теоретическая, a priori, как конечная последовательность случайных величин или векторов $(\xi_1, \xi_2, \ldots \xi_n)$;
- выборка статистическая, a posteriori, как конечная последовательность значений $(x_1, x_2, \ldots x_n)$, принятых в результате эксперимента элементами теоретической выборки.

В настоящем разделе мы будем иметь в виду выборку статистическую, каждый член которой - скаляр (одномерную статистическую выборку). Для сокращения речи будем говорить «выборка».

Исходным материалом любого статистического рассуждения есть выборка

$$x_1, x_2, \ldots x_n, \tag{18.1}$$

число n называется ее *объемом*. Для получения содержательных и надежных выводов желательно объем выборки делать как можно большим. При этом получаемые числовые данные могут стать труднообозримыми. С целью облегчения усвоения информации, содержащей в выборке, ее подвергают различным преобразованиям.

Целесообразно прежде всего выборочные данные *ранжировать*, располагая их в порядке возрастания и приходя в результате к *вариационному* ряду:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}, \tag{18.2}$$

где $x_{(1)}$ - наименьшее, $x_{(n)}$ - наибольшее из чисел $(x_1, x_2, \ldots x_n)$, $x_{(i)}$ - i-е по величине. При этом может случиться, что количество различных членов в вариационном ряду окажется меньше объема исходной выборки (равенство членов выборки не исключается). В этой ситуации данные представляют в виде *группированного* вариационного ряда таблицей

$x_{(i)}$	$x_{(1)}$	$x_{(2)}$	. . .	$x_{(k)}$
n_i	n_1	n_2	. . .	n_k

, (18.3)

где $x_1, x_2, \ldots x_k$ - различные члены

вариационного ряда ($x_{(1)} \le x_{(2),} \le \ldots \le x_{(k)}$), $n_1, \ldots, n_k$ - количество повторений чисел $x_1, x_{2,} \ldots x_k$ в вариационном ряду - *частоты*, $\sum_{i=1}^{k} n_i = n$.

Пример 18.1. Страховая компания, занимающаяся обязательным страхованием гражданской ответственности транспортных средств, имеет в своем распоряжении данные о количестве дорожно-транспортных происшествий на протяжении 56 равных непрерывающихся промежутков времени

5, 1, 4, 5, 4, 3, 5, 5, 2, 5, 5, 6, 4, 3, 1, 5, 2, 5, 5, 5, 3, 3, 3, 6, 6, 5, 6, 5, 3, 4, 5, 4, 6, 6, 5, 2, 1, 5, 4, 5, 5, 3, 6, 4, 5, 5, 4, 3, 5, 5, 5, 4, 5, 4, 2, 4, 5, 6, 1, 5.

Составим вариационный ряд

$$\underbrace{1111}_{4}\underbrace{1}\underbrace{2222}_{4}\underbrace{333333}\underbrace{33}\underbrace{4444444444}_{10}\underbrace{4}\underbrace{55555555}_{22}\underbrace{5\ldots5}\underbrace{55}\underbrace{66666666}_{8}$$

и группированный вариационный ряд

Таблица 18.1

(Данные о числе ДТП)

$x_{(i)}$	1	2	3	4	5	6	
n_i	4	4	8	10	22	8	$\sum$ 56

☑ На практике эти две операции при обработке информации вручную объединяют в одну, составляя таблицу

$x_{(i)}$	1	2	3	4	5	6	
Счет	□	□	⊠Γ	⊠□	⊠⊠ ⊠□	⊠Γ	
n_i	4	4	8	10	22	8	$\sum$ 56

В случае, если объем выборки совпадает с объемом вариационного ряда (n = k), группировка данных может производиться не по самим значениям вариант, а группировкой их в интервалы. Для этого

- а) определяют *размах* выборки: $x_{(n)} - x_{(1)} = x_{max} - x_{min}$;
- б) определяют *шаг* выборки, пользуясь формулой Стэрджеса

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322 \cdot \lg n}; \qquad (18.4)$$

- определяют начало первого интервала, a_1, полагая $a_1 = (x_{min} - h/2)$, $a_2 = a_1 + h/2$, $a_3 = a_2 + h$, ... и т.д. до тех пор, пока a_k впервые не станет больше x_{max}. Далее считают $n_1, \ldots, n_k$ - количество $x_1, x_2, \ldots x_k$, попавших в i интервал.

Впрочем, эти рекомендации следует применять, сообразуясь с удобствами вычислений и дальнейшего графического представления данных.

Пример 18.2. Руководитель офиса заинтересовался длительностью междугородных телефонных разговоров, имевших место в течение одной недели. Ему представили данные, зафиксированные номеронабирателем (в минутах)

11,8	3,6	16,6	13,5	4,8	11,2	10,4	7,2	5,5	14,5
8,3	8,9	9,1	7,7	2,3	8,5	15,9	18,7	11,7	6,2
12,1	6,1	10,2	8,0	11,4	6,8	9,6	19,5	15,3	12,3

Представим эти данные в виде интервального вариационного ряда, определив:

- а) размах выборки $x_{max} - x_{min} = 19,5 - 2,3 = 17,2$;
- б) шаг группировки $h = \dfrac{17,2}{1 + 3,322 \cdot \lg 30} \cong 2,912$ (полагаем h = 3);
- в) начало первого интервала $a_1 = 2,3 - 1,5 = 0,8$ (полагаем $a_1 = 1$).

Строим интервальный вариационный ряд

Интервалы	(1;4)	(4;7)	(7;10)	(10;13)	(13;16)	(16;19)	(19;22)
Счет	Г	⧅	⧅Г	⧅Г	□	Г	I
$\mathbf{n_i}$	2	5	8	8	4	2	1

Те же данные, сгруппированные с $a_1 = 2$, приведут нас к иному интервальному вариационному ряду

Таблица 18.2

(Данные о длительности разговоров)

Интервалы	(2;5)	(5;8)	(8;11)	(11;14)	(14;17)	(17;20)
$\mathbf{n_i}$	3	6	8	7	4	2

Для оценки количества интервалов группировки в зависимости от объема выборки можно пользоваться следующей таблицей:

Объем выборки, **n**	менее 50	50 - 200	200 - 500	500 - 1000
Количество классов, **k**	5 -7	7 - 9	9 - 10	10 - 11

Числа n_i в табл. 18.1 или аналогичные им в интервальном вариационном ряду (табл. 18.2) носят название *частот* соответствующих членов или интервалов вариационного ряда.

Числа

$$W_i = \frac{n_i}{n}, \quad i = 1,2,\ldots,k \tag{18.5}$$

называют *относительными частотами* или *частостями*,

числа

$$n_i^* = \sum_{j \le i} n_j, \quad i = \overline{1,k}; \tag{18.6}$$

и

$$w_i^* = \sum_{j \le i} w_j, \quad i = \overline{1,k}. \tag{18.7}$$

носят соответственно название *накопленных частот* и *накопленных частостей*. ◆

Вычисление чисел n_i, w_i, n_i^*, w_i^* предшествует графическому представлению данных.

§19. Графическое представление вариационных рядов

Пример 19.1. По данным примеров 18.1 и 18.2 построим таблицы, содержащие частоты и частости.

Таблица 19.1

(Данные о числе ДТП)

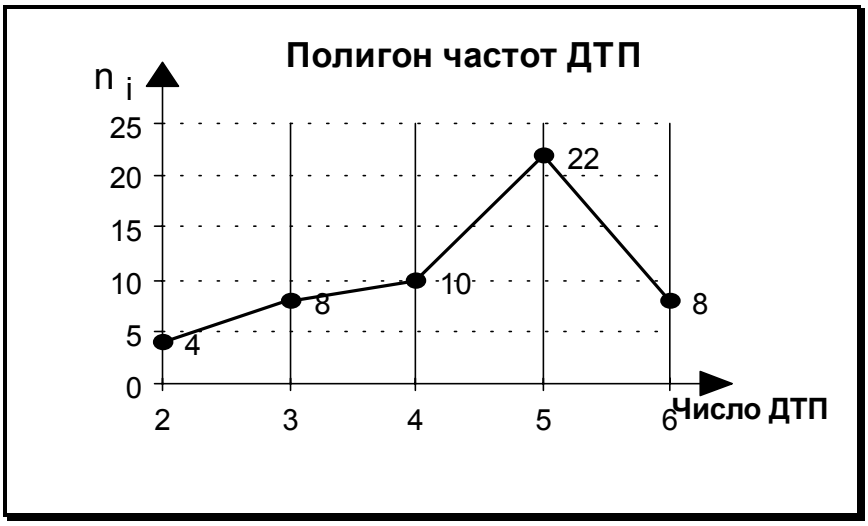
i	x_i	n_i	n_i^*	w_i	w_i^*
1	1	4	4	0,07	0,07
2	2	4	8	0,07	0,14
3	3	8	16	0,14	0,28
4	4	10	26	0,19	0,47
5	5	22	48	0,39	0,86
6	6	8	56	0,14	1,00
Сумма		56		1,00	

Таблица 19.2

(Данные о длительности разговоров)

i	Интервалы	n_i	n_i^*	w_i	w_i^*
1	2; 5	3	3	0,10	0,10
2	5; 8	6	9	0,20	0,3
3	8; 11	8	17	0,27	0,57
4	11; 14	7	24	0,23	0,80
5	14; 17	4	28	0,13	0,93
6	17; 20	2	30	0,07	1,00
Сумма		30		1,00	

Полигоном частот (частостей) называют кусочно-линейную ломаную с вершинами в точках (x_i^*, n_i) $((x_i^*, w_i))$, где x_i^* - середины интервалов в случае интервального вариационного ряда.



На рис. (19.1а), (19.1б), (19.2а), (19.2б) показаны полигоны и гистограммы частот и частостей, построенные по данным табл. (19.1) и (19.2)

Рис. 19.1а

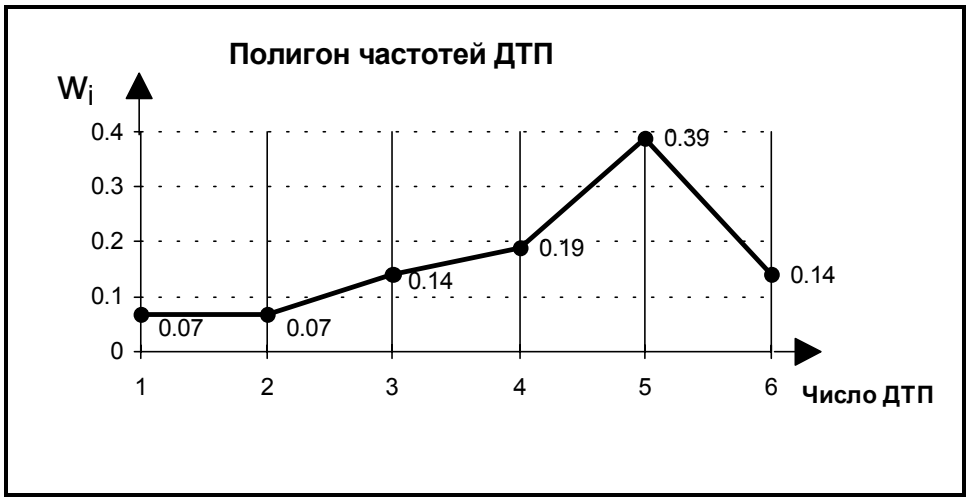


Рис. 19.1.б

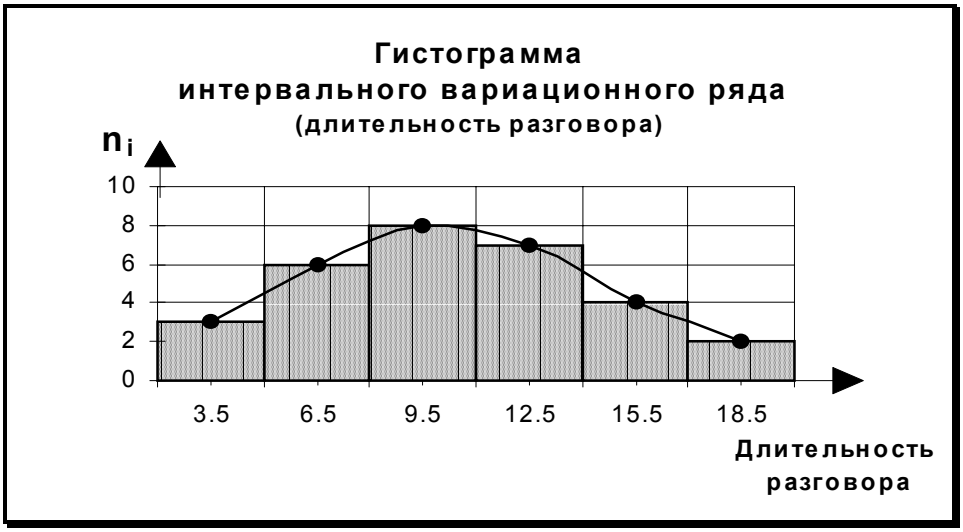


Рис.19.2.а

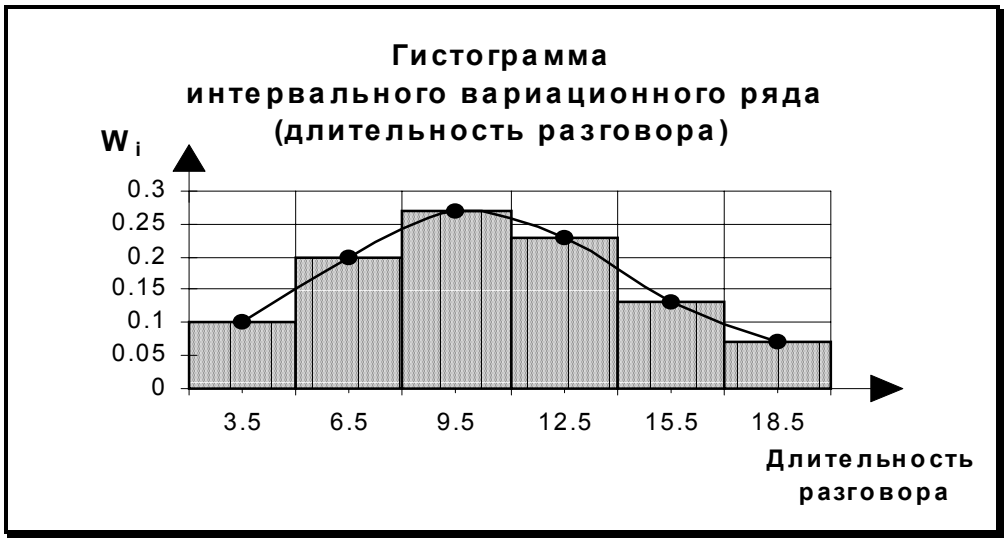


Рис.19.2.б

Огивой или *кумулятивной кривой накопленных частостей* называют ломаную, построенную по точкам (x_i, w_i^*), соответственно по точкам (x_i^*, w_i^*).

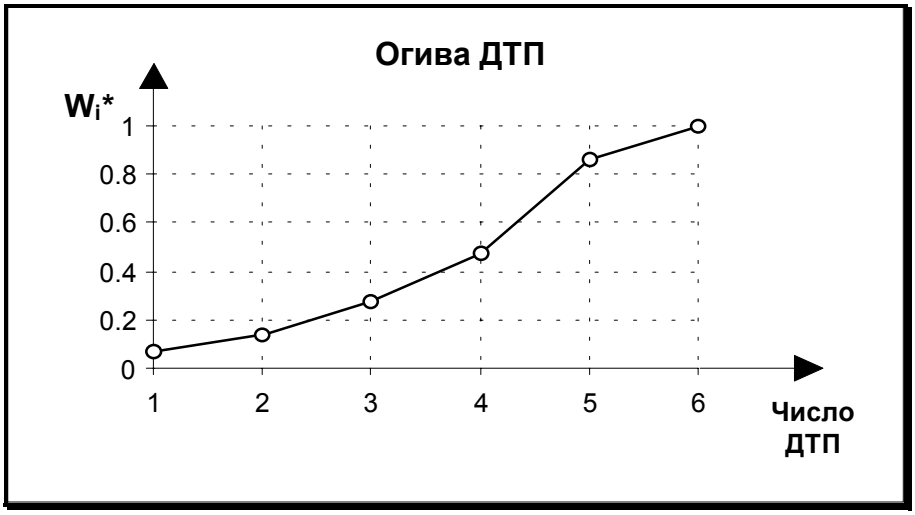


Рис. 19.3

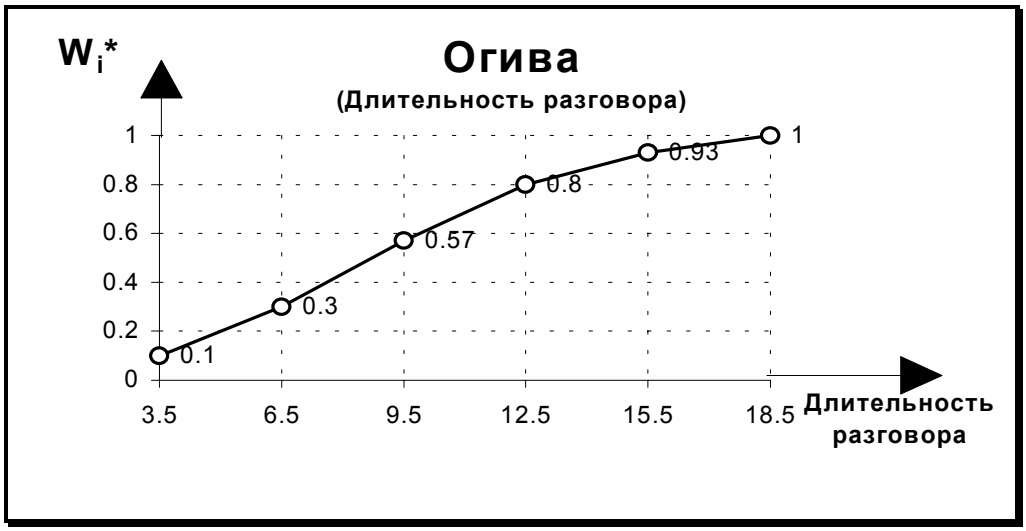


Рис. 19.4

На рис. 19.3 и 19.4 показаны огивы, построенные по табл. 19.1 и 19.2 соответственно. ◆

Первоначальное табличное или графическое представление выборочных данных сопровождают вычислением некоторых числовых характеристик.

§20. Числовые характеристики вариационных рядов

Для единообразия формул будем обозначать x_i выборочное значение независимо от того, в каком из вариационных рядов оно находится.

- *Начальным выборочным моментом* порядка r называют число

$$m_r = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i^r, \quad r = 1,2,\ldots.$$

- *Центральным выборочным моментом* порядка r называют число

$$d_r = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - m_1)^r, \quad r = 1,2,\ldots .$$

Моменты порядков $r = 1$ и $r = 2$ наиболее употребительны и носят специальные названия и обозначаются особо.

Момент m_1 (порядка 1) обозначается $\bar{x}$ и носит название *выборочного среднего*

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot n_i .$$

Свойства выборочного среднего аналогичны свойствам математического ожидания, одно из них выглядит так:

$$\overline{ax + b} = a \cdot \bar{x} + b. \qquad (20.1)$$

В самом деле, значения вариационного ряда для выборки из генеральной совокупности $a\xi + b$ будут таковы: $a \cdot \overline{x_i} + b$, $i = \overline{1,k}$; частоты при этом остаются теми же. Поэтому

$$\overline{ax + b} = \frac{1}{n} \cdot \sum_{i=1}^{k} (ax_i + b) \cdot n_i = a \cdot \frac{1}{n} \sum_{i=1}^{k} x_i \cdot n_i + b \cdot \frac{1}{n} \cdot \sum_{i=1}^{k} n_i = a \cdot \bar{x} + b .$$

Центральный момент порядка 1 всегда равен 0. Центральный момент порядка 2 называют *выборочной дисперсией* и обозначают S_x^2:

$$S_x^2 = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i .$$

Свойства выборочной дисперсии аналогичны свойствам дисперсии.

Например,

$$S_{ax+b}^2 = a^2 \cdot S_x^2 . \qquad (20.2)$$

В самом деле,

$$S_{ax+b}^2 = \frac{1}{n} \cdot \sum_{i=1}^{k} (ax_i + b - \overline{(ax + b)})^2 \cdot n_i = \frac{1}{n} \cdot \sum_{i=1}^{k} (ax_i + b - a\bar{x} - b)^2 \cdot n_i =$$

$$= \frac{a^2}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i = a^2 \cdot S_x^2 .$$

При вычислении на практике удобно пользоваться такой формулой:

$$S_x^2 = \frac{1}{n} \cdot \sum_{i=1}^{k} x_ø^2 \cdot n_i - (\bar{x})^2 = m_2 - (m_1)^2 \,. \qquad (20.3)$$

Закончим этот параграф вычислением числовых характеристик количества дорожно-транспортных происшествий, сведя все вычисления в таблицу.

Таблица 20.1.

(Числовые характеристики количества ДТП)

i	x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
1	1	4	4	1	4
2	2	4	8	4	16
3	3	8	24	9	72
4	4	10	40	16	160
5	5	22	110	25	550
6	6	8	48	36	288
Суммы		56	234		1090

Теперь вычисляем

$$\bar{x} = \frac{234}{56} \cong 4.18; \qquad S_x^2 = \frac{1090}{56} - \left(\frac{234}{56}\right)^2 = \frac{6284}{3136} \cong 2.00.$$

◆

Соответствующие вычисления для данных из примера 18.2 таковы:

Таблица 20.2.

(Числовые характеристики длительности разговоров)

i	x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
1	3,5	3	10.5	12.25	36.75
2	6,5	6	39.0	42.25	253.50
3	9,5	8	76.0	90.25	722.00
4	12,5	7	87.5	156.25	1093.75
5	15,5	4	62.0	240.25	961.00
6	18,5	2	37.0	342.25	684.50
Суммы		30	312		3751.5

Теперь получаем

$$\bar{x} = \frac{312}{30} \cong 10{,}4; \qquad S_x^2 = \frac{3751{,}5}{30} - \left(\frac{312}{30}\right)^2 \cong 16{,}89.$$

Здесь для вычисления $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot n_i$ в качестве $x_1, x_2, \ldots x_n$ используют представителя интервала, например, среднее значение интервала.

◆

§21. Статистические оценки неизвестных параметров распределения

Как уже говорилось, изучая генеральную совокупность, мы на самом деле изучаем некоторую случайную величину ξ. Последняя же становится известной, если известно ее распределение (плотность или функция распределения). В типичных задачах вид функции распределения ξ известен с точностью до некоторых параметров.

Например:

- в теории измерений считается, что ошибки измерений подчинены нормальному закону с плотностью

$$f(x,a,\sigma^2)=\frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\},\quad x\in R^1$$

и задача состоит в определении a либо σ, либо обоих параметров одновременно.

- Количество несчастных случаев часто предполагают распределенным по закону Пуассона

$$P_k(\lambda)=\frac{\lambda^k}{k!}\cdot e^{-\lambda},\quad k=0,1,2,3,...$$

и задача состоит в определении параметра λ.

- В теории надежности предполагают, что длительность безотказной работы прибора подчинена показательному закону распределения с параметром μ:

$$f(x,\mu)=\mu\cdot e^{-\mu\cdot x},\quad x>0,$$

и снова возникает задача об определении одного параметра.

Вся информация, которой мы располагаем при решении указанных задач, содержится в выборке $(\xi_1,\xi_2,...\xi_n)$.

Определение 21.1. Оценкой неизвестного параметра Θ называют любую (борелевскую) функцию n переменных от выборки:

$$\tilde{\theta}_n=f(\xi_1,\xi_2,...,\xi_n).$$

Поскольку оценка является функцией от случайного вектора (выборки), она сама является случайной величиной, распределение которой зависит от числа наблюдений n и оцениваемого параметра Θ.

§22. Основные свойства точечных оценок

Для того чтобы оценка $\tilde{\theta}_n$ имела практическую ценность, она должна обладать следующими свойствами.

- 1. Оценка $\tilde{\theta}_n$ параметра θ называется *несмещенной*, если ее математическое ожидание равно оцениваемому параметру θ, т.е.

$$M\tilde{\theta}_n = \theta. \qquad (22.1)$$

Если равенство (22.1) не выполняется, то оценка $\tilde{\theta}_n$ может либо завышать значение θ ($M\tilde{\theta}_n > \theta$), либо занижать его ($M\tilde{\theta}_n < \theta$). Естественно в качестве приближенного неизвестного параметра брать несмещенные оценки для того, чтобы не делать систематической ошибки в сторону завышения или занижения.

- 2. Оценка $\tilde{\theta}_n$ параметра θ называется *состоятельной*, если она подчиняется закону больших чисел, т.е. *сходится по вероятности* к оцениваемому параметру при неограниченном возрастании числа опытов (наблюдений) и, следовательно, выполняется следующее равенство:

$$\lim_{n\to\infty} P\{|\tilde{\theta}_n - \theta| < \varepsilon\} = 1, \qquad (22.2)$$

где $\varepsilon > 0$ сколько угодно малое число.

Для выполнения (22.2) достаточно, чтобы дисперсия оценки стремилась к нулю при $n \to \infty$, т.е.

$$\lim_{n\to\infty} D(\tilde{\theta}_n) = 0 \qquad (22.3)$$

и кроме того, чтобы оценка была несмещенной. От формулы (22.3) легко перейти к (22.2), если воспользоваться неравенством Чебышева.

Итак, состоятельность оценки означает, что при достаточно большом количестве опытов и со сколько угодно большой достоверностью отклонение оценки от истинного значения параметра меньше любой наперед заданной величины. Этим оправдано увеличение объема выборки.

Так как $\tilde{\theta}_n$ - случайная величина, значение которой изменяется от выборки к выборке, то меру ее рассеивания около математического ожидания θ будем характеризовать дисперсией $D\tilde{\theta}_n$. Пусть $\tilde{\alpha}_n^1$ и $\tilde{\alpha}_n^2$ - две несмещенные оценки параметра θ, т.е. $M\tilde{\alpha}_n^1 = \theta$ и $M\tilde{\alpha}_n^2 = \theta$, соответственно $D\tilde{\alpha}_n^1$ и $D\tilde{\alpha}_n^2$ и, если $D\tilde{\alpha}_n^1 < D\tilde{\alpha}_n^2$, то в качестве оценки принимают $\tilde{\alpha}_n^1$.

- 3. Несмещенная оценка $\tilde{\theta}_n$, которая имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра θ, вычисленных по выборкам одного и того же объема, называется *эффективной* оценкой.

На практике при оценке параметров не всегда удается удовлетворить одновременно требованиям 1, 2, 3. Однако выбору оценки всегда должно предшествовать ее критическое рассмотрение со всех точек зрения. При выборке практических методов обработки опытных данных необходимо руководствоваться сформулированными свойствами оценок.

§23. Оценка математического ожидания и дисперсии по выборке

Наиболее важными характеристиками случайной величины являются математическое ожидание и дисперсия. Рассмотрим вопрос о том, какие выборочные характеристики лучше всего оценивают математическое ожидание и дисперсию в смысле несмещенности, эффективности и состоятельности.

Теорема 23.1. Арифметическая средняя $\bar{x}$, вычисленная по n независимым наблюдениям над случайной величиной ξ, которая имеет математическое ожидание $M\xi = \mu$, является несмещенной оценкой этого параметра.

Д о к а з а т е л ь с т в о.

Пусть $\xi_1, \xi_2, \ldots \xi_n$ - n независимых наблюдений над случайной величиной ξ. По условию $M\xi = \mu$, а т.к. ξ_i $(i = \overline{1:n})$ являются случайными величинами и имеют тот же закон распределения, то тогда $M\xi_i = \mu$ $(i = \overline{1:n})$. По определению средняя арифметическая

$$\bar{\xi} = \frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} \xi_i . \qquad (23.1)$$

Рассмотрим математическое ожидание средней арифметической. Используя свойство математического ожидания, имеем:

$$M\bar{\xi} = M\left(\frac{1}{n} \cdot \sum_{i=1}^{n} \xi_i\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} M\xi_i = \frac{1}{n} \cdot n \cdot \mu = \mu ,$$

т.е. $M\bar{\xi} = \mu$. В силу (22.1) $\bar{x}$ является несмещенной оценкой. ■

Теорема 23.2. Арифметическая средняя $\bar{\xi}$, вычисленная по n независимым наблюдениям над случайной величиной ξ, которая имеет $M\xi = \mu$ и $D\xi = \sigma^2$, является состоятельной оценкой этого параметра.

Д о к а з а т е л ь с т в о.

Пусть $\xi_1, \xi_2, \ldots \xi_n$ - n независимых наблюдений над случайной величиной ξ. Тогда в силу теоремы 23.1 имеем $M\xi = M\bar{\xi} = \mu$.

Для средней арифметической $\bar{\xi}$ запишем неравенство Чебышева:

$$P\{\left|\bar{\xi} - \mu\right| \geq \varepsilon\} \leq \frac{D\bar{\xi}}{\varepsilon^2} .$$

Используя свойства дисперсии 4,5 и (23.1), имеем:

$$D\bar{\xi} = \frac{1}{n^2} \cdot \sum_{i=1}^{n} D\xi_i = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n^2} \cdot n ,$$

т.к. по условию теоремы $Dx_i = D\xi = \sigma^2$.
Следовательно,

$$D\bar{\xi} = \frac{1}{n^2}\cdot\sum_{i=1}^{n}\sigma^2 = \frac{n\cdot\sigma^2}{n^2} = \frac{\sigma^2}{n}\ . \qquad (23.2)$$

Итак, дисперсия средней арифметической в n раз меньше дисперсии случайной величины ξ. Тогда

$$\lim_{n\to\infty}\frac{D\bar{\xi}}{\varepsilon^2} = \lim_{n\to\infty}\frac{\sigma^2}{n\cdot\varepsilon^2} = 0,$$

поэтому

$$\lim_{n\to\infty} P\{|\bar{\xi}-\mu| < \varepsilon\} = 1,$$

а это значит, что $\bar{\xi}$ является состоятельной оценкой.

З а м е ч а н и е : 1. Примем без доказательства весьма важный для практики результат. Если $\xi \in N(a, \sigma)$, то несмещенная оценка $\bar{\xi}$ математического ожидания **a** имеет минимальную дисперсию, равную $\frac{\sigma^2}{n}$, поэтому $\bar{\xi}$ является эффективной оценкой параметра a. ■

Перейдем к оценке для дисперсии и проверим ее на состоятельность и несмещенность.

Теорема 23.3. Если случайная выборка состоит из n независимых наблюдений над случайной величиной ξ с $M\xi = \mu$ и $D\xi = \sigma^2$, то выборочная дисперсия

$$S^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2 \qquad (23.3)$$

не является несмещенной оценкой $D\xi$ - генеральной дисперсии.

Д о к а з а т е л ь с т в о.

Пусть $\xi_1, \xi_2, \ldots \xi_n$ - n независимых наблюдений над случайной величиной ξ. По условию $M\xi_i = M\xi = \mu$ и $D\xi_i = D\xi = \sigma^2$ для всех $i = \overline{1:n}$. Преобразуем формулу (23.3) выборочной дисперсии:

$$S^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}(\xi_i - \mu + \mu - \bar{\xi})^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}\left[(\xi_i - \mu) - (\bar{\xi} - \mu)\right]^2 =$$

$$= \frac{1}{n}\cdot\sum_{i=1}^{n}\left[(\xi_i - \mu)^2 - 2(\xi_i - \mu)\cdot(\bar{\xi} - \mu) + (\bar{\xi} - \mu)^2\right] =$$

$$= \frac{1}{n}\cdot\sum_{i=1}^{n}(\xi_i - \mu)^2 - \frac{2}{n}\cdot(\bar{\xi} - \mu)\cdot\sum_{i=1}^{n}(\xi_i - \mu) + \frac{n}{n}\cdot(\bar{\xi} - \mu)^2.$$

Упростим выражение

$$\frac{2}{n}\cdot(\bar{\xi} - \mu)\cdot\sum_{i=1}^{n}(\xi_i - \mu).$$

Принимая во внимание (23.1), откуда

$$n \cdot \bar{\xi} = \sum_{i=1}^{n} \xi_i$$

можно записать

$$\frac{1}{n} \cdot 2 \cdot (\bar{\xi} - \mu) \cdot \sum_{i=1}^{n} (\xi_i - \mu) = \frac{1}{n} \cdot (\bar{\xi} - \mu) \cdot (\sum_{i=1}^{n} \xi_i - n \cdot \mu) = \frac{1}{n} \cdot 2 \cdot (\bar{\xi} - \mu) \cdot (n \cdot \bar{\xi} - n \cdot \mu) = 2 \cdot (\bar{\xi} - \mu)^2$$

Тогда

$$S^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\xi_i - \mu)^2 - (\bar{\xi} - \mu)^2.$$

Теперь рассмотрим MS^2 - математическое ожидание выборочной дисперсии:

$$MS^2 = \frac{1}{n} \cdot M\left[\sum_{i=1}^{n} (\xi_i - \mu)^2\right] - M(\bar{\xi} - \mu)^2.$$

Используя определение дисперсии, получаем:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} M(\xi_i - \mu)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} D\xi_i = \frac{1}{n} \cdot n \cdot \sigma^2 = \sigma^2$$

и $\quad M(\bar{\xi} - \mu)^2 = D\bar{\xi} = \frac{1}{n} \cdot \sigma^2$ в силу (23.2), следовательно,

$$MS^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \cdot \sigma^2, \qquad (23.4)$$

т.е. выборочная дисперсия является смещенной оценкой дисперсии генеральной совокупности.

З а м е ч а н и е 2. Оценку (23.4) можно исправить так, чтобы она стала несмещенной

$$\hat{S}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2. \qquad (23.5)$$

Обычно оценку $\hat{S}^2$ называют *исправленной выборочной дисперсией*. Действительно,

$$\hat{S}^2 = \frac{n}{n-1} \cdot S^2,$$

тогда

$$M\hat{S}^2 = \frac{n}{n-1} \cdot MS^2 = \frac{n}{n-1} \cdot \frac{n-1}{n} \cdot \sigma^2 = \sigma^2.$$

Дробь $\frac{n}{n-1}$ называют *поправкой Бесселя*. При малых n поправка Бесселя значительно отличается от 1. При n > 50 практически нет разницы между S^2 и $\hat{S}^2$.

З а м е ч а н и е 3. Можно показать, что оценки S^2 и $\hat{S}^2$ являются состоятельными и не являются эффективными.

Несмещенной, состоятельной и эффективной оценкой σ^2 является оценка

$$S_*^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\xi_i - \mu)^2 \qquad (23.6)$$

в случае, когда математическое ожидание μ известно.

§24. Доверительные интервалы

Изучавшиеся ранее оценки неизвестного параметра являются *точечными:* мы старались судить о значении неизвестного числа или вектора θ по значению оценки $\tilde{\theta}_n = f(\xi_1,\xi_2,...,\xi_n)$, принятом ею, как только известна статистическая выборка $(x_1,x_2,...,x_n)$. Однако, поскольку оценка сама является случайной величиной, её выборочное значение заведомо не совпадает с константой θ. Имея в виду это обстоятельство, предпочтительнее стремиться указывать не точное значение оцениваемого параметра, а некоторый интервал, содержащий в себе значение параметра. Границы такого интервала должны определяться доступной нам информацией - выборкой из генеральной совокупности, то есть они сами случайны, и поэтому есть смысл говорить о вероятности того, что значение параметра находится внутри интервала.

Определение 24.1. Пусть генеральная совокупность описывается случайной величиной ξ, распределение которой зависит от скалярного параметра θ. Пусть, далее, $\theta_1 = \theta_1(\xi_1,\xi_2,...,\xi_n)$ и $\theta_2 = \theta_2(\xi_1,\xi_2,...,\xi_n)$ две функции выборки такие, что всегда $\theta_1 \le \theta_2$ и

$$P\{\theta_1(\xi_1,\xi_2,...,\xi_n) < \theta < \theta_2(\xi_1,\xi_2,...,\xi_n)\} = \beta.$$

(θ_1,θ_2) со случайными границами называют *доверительным* интервалом для неизвестного параметра θ с доверительной вероятностью β.

Число $\alpha = 1-\beta$ называют *уровнем значимости* интервала.

Стараясь иметь как можно более достоверные выводы, границы доверительного интервала выбирают таким образом, чтобы доверительная вероятность β была как можно ближе к 1.

Схематически процесс построения доверительного интервала можно описать следующим образом.

Пусть $\tilde{\theta}_n$ - несмещенная оценка параметра θ.

Выберем доверительную вероятность β. Значение выражения «β как можно ближе к 1» относительно, оно находится вне границ математики и определяется лицом, производящим статистические исследования. Обычно выбирают β равным 0,9; 0,95; 0,99.

Пусть, далее, можно найти такое число $\varepsilon > 0$, что

$$P\{|\tilde{\theta}_n - \theta| < \varepsilon\} = \beta. \qquad (24.1)$$

Записав (24.1) в виде

$$P\{\tilde{\theta}_n - \varepsilon < \theta < \tilde{\theta}_n + \varepsilon\} = \beta,$$

видим, что интервал $(\tilde{\theta}_n - \varepsilon, \tilde{\theta}_n + \varepsilon)$ является доверительным интервалом для параметра θ с уровнем значимости $\alpha = 1-\beta$.

Практически вопрос о построении доверительного интервала связан с возможностью нахождения распределения оценки $\tilde{\theta}_n$, а это, в свою очередь, зависит от распределения генеральной совокупности.

Пример 24.1. *Построение доверительного интервала для математического ожидания нормальной генеральной совокупности при известной дисперсии.*

Пусть генеральная совокупность ξ распределена по нормальному закону с параметрами (θ,σ^2), где σ^2 (дисперсия) известно. Мы уже знаем, что наилучшей в смысле несмещенности, состоятельности и эффективности оценкой неизвестного математического ожидания θ нормального закона является выборочное среднее

$$\bar{\xi} = \frac{1}{n} \cdot \sum_{i=1}^{n} \xi_i .$$

В продвинутом курсе теории вероятностей доказывается, что нормальное распределение обладает свойством *устойчивости* : если независимые случайные величины ξ, η распределены нормально с параметрами (a_1,σ_1^2) и (a_2,σ_2^2) соответственно, то их сумма $\xi + \eta$ распределена нормально с параметрами $(a_1 + a_2, \sigma_1^2 + \sigma_2^2)$.

Используя это утверждение в нашем случае, заключаем, что $\bar{\xi}$ распределена нормально с параметрами $(\theta, \frac{\sigma^2}{n})$, а нормированное выборочное среднее $\mu = \frac{(\bar{\xi} - \theta)\sqrt{n}}{\sigma}$ подчинено нормальному закону с параметрами $(0,1)$.

Это означает, что

$$P\left\{\left|\frac{(\bar{\xi} - \theta)\sqrt{n}}{\sigma}\right| < Z\right\} = 2\Phi(z),$$

где $\Phi(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^z e^{-\frac{x^2}{2}} dx$.

Функция $\Phi(z)$ нам уже встречалась, её значения табулированы.

Выберем теперь доверительную вероятность β и обозначим $z_{\frac{\beta}{2}}$ корень уравнения $\Phi(z_{\frac{\beta}{2}}) = \beta/2$.

Рассмотрим равенства

$$P\left\{\left|\frac{(\bar{\xi}-\theta)\sqrt{n}}{\sigma}\right| < Z_{\frac{\beta}{2}}\right\} = 2\Phi(z_{\frac{\beta}{2}}) = \beta = P\left\{\left|(\bar{\xi}-\theta)\right| < \frac{\sigma}{\sqrt{n}} Z_{\frac{\beta}{2}}\right\} = P\{\bar{\xi} - \frac{\sigma}{\sqrt{n}} z_{\frac{\beta}{2}} < \theta < \bar{\xi} + \frac{\sigma}{\sqrt{n}} z_{\frac{\beta}{2}}\},$$

которые свидетельствуют о том, что интервал

$$(\bar{\xi} - \frac{\sigma}{\sqrt{n}} z_{\frac{\beta}{2}}, \bar{\xi} + \frac{\sigma}{\sqrt{n}} z_{\frac{\beta}{2}})$$

является доверительным для параметра θ с доверительной вероятностью β (и уровнем значимости $\alpha = 1 - \beta$).

Приведем часть из таблицы значений $z_{\frac{\beta}{2}}$ (прил. 2) для некоторых наиболее употребительных значений β.

Таблица 24.1

Зависимость $z_{\frac{\beta}{2}}$ от доверительной вероятности

β	0,9	0,925	0,95	0,99
$z_{\frac{\beta}{2}}$	1,65	1,78	1,96	2,89

Обозначим $\Delta = Z_{\frac{\beta}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ половину ширины доверительного интервала.

Замечаем, что:

1) при фиксированной доверительной вероятности β ширина доверительного интервала уменьшается с ростом числа наблюдений n как величина порядка $\frac{1}{\sqrt{n}}$ (при увеличении, например, числа наблюдений в 100 раз ширина интервала уменьшится в 10 раз);

2) поскольку $\Phi(z)$ возрастает с ростом z, то увеличение доверительной вероятности, при всех прочих постоянных параметрах, приводит к расширению доверительного интервала.

Пример 24.2. Желая узнать, сколько часов в неделю дети проводят у телевизора, социологическая служба обследовала 100 учеников некого города, в результате чего оказалось, что в среднем это число равно $\bar{x} = 27.5$. Из прошлой практики известно, что стандартное отклонение ($\sigma = \sqrt{D\xi}$) генеральной совокупности равно 6 (часов). Найдем доверительный интервал с доверительной вероятностью 0,95 для числа часов в неделю, проводимых ребенком у телевизора.

Поскольку $\beta = 0{,}95$, из табл. 24.1 находим $z_{0{,}475} = 0{,}96$, и границы интервала доверия будут такими:

$$\bar{x} \pm z_{0{,}475} \cdot \frac{\sigma}{\sqrt{n}} = 27{,}5 \pm 1{,}.96 \cdot \frac{6{,}0}{\sqrt{100}} = 27{,}5 \pm 1{,}18,$$

интервал доверия имеет вид (26.32; 28.68).

Теперь поставим вопрос иначе: сколько детей надо обследовать с тем, чтобы среднее число часов в неделю, проводимых ребенком у телевизора, отклонилось от его оценки не более чем на 0,5 ч. с вероятностью 0,95?

В такой постановке речь идет о нахождении числа n таким, чтобы выполнялось равенство

$$P\left\{\left|\bar{\theta}_n - \theta\right| \le \frac{1}{2}\right\} = P\left\{\left|\frac{(\bar{\xi} - \theta)\sqrt{n}}{\sigma}\right| \le \frac{\sqrt{n}}{2\sigma}\right\} = 0{,}95\ ,$$

откуда $\frac{\sqrt{n}}{2\sigma} = Z_{0{,}475}$ или $n = (2\sigma Z_{0.475})^2$.

В условиях примера $n = (2 \cdot 6 \cdot 1{,}96)^2 \cong 553$.

Разумеется, при больших значениях n ширина доверительного интервала уменьшится.

Заметим, что по сравнению с первоначальной задачей ширина интервала уменьшилась в 1,18/0,5 = 2,36 раз, количество необходимых испытаний увеличилось в $(2{,}36)^2 = 5{,}57$ раз (553 отличается в третьем знаке от $100 \cdot 5{,}57$). ◆

Пример 24.3. *Построение доверительного интервала для математического ожидания нормальной генеральной совокупности при неизвестной дисперсии.*

Снова рассмотрим генеральную совокупность ξ, распределенную нормально с параметрами (θ, σ^2), однако теперь считаем дисперсию σ^2 неизвестной.

Обозначим $\hat{S}$ стандартное выборочное квадратичное отклонение

$$\hat{S} = \sqrt{\hat{S}^2} = \sqrt{\frac{1}{n-1} \cdot \sum_{i-1}^{n} (\xi_i - \bar{\xi})^2}\ .$$

В курсах теории вероятностей доказывается, что случайная величина

$$t = \frac{\bar{\xi} - \theta}{\hat{S}} \cdot \sqrt{n}$$

подчиняется так называемому закону распределения Стьюдента с n - 1 степенью свободы и её плотность имеет вид

$$t_{n-1}(x) = K_n \cdot \left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}},$$

где K_n некоторая нормирующая константа.

Созданы таблицы , дающие возможность вычислять вероятности вида

$$\alpha = P\left\{\left|t_{n-1}\right| > Z_{n-1,\alpha}\right\} = 2\int_{z_{n-1,\alpha}}^{\infty} t_{n-1}(x)dx$$

(см. прил. 4).

Ввиду вышесказанного, получаем равенства:

$$P\left\{\left|\frac{(\overline{\xi} - \theta)\cdot\sqrt{n}}{\hat{S}}\right| < Z\right\} = P\left\{\left|\overline{\xi} - \theta\right| < Z\cdot\frac{\hat{S}}{\sqrt{n}}\right\} = 2\int_{0}^{z_{n-1,\alpha}} t_{n-1}(x)dx,$$

из которых видно, что выбрав Z как корень уравнения

$$2\int_{z}^{\infty} t_{n-1}(x)dx = P\left\{\left|t_{n-1}\right| > Z\right\} = 1-\beta$$

(обозначим этот корень $Z_{n-1,1-\beta} = Z_{n-1,\alpha}$), приходим к доверительному интервалу для θ вида

$$(\overline{\xi} - \frac{Z_{n-1,\alpha}}{\sqrt{n}}\cdot\hat{S} < \theta < \overline{\xi} + \frac{Z_{n-1,\alpha}}{\sqrt{n}}\cdot\hat{S}).$$ ◆

Пример 24.4. Рассмотрим вопрос о построении доверительного интервала для неизвестного количества времени в течение недели, проводимого ребенком у экрана телевизора, сохранив все данные примера 24.2, считая теперь, что 6ч. есть оценка выборочного среднеквадратического отклонения, $\hat{S} = 6$.

По таблице распределения Стьюдента (см. приложение 4) находим $z_{99;0,995} = 2{,}626$, границы интервала будут

$$(27{,}5 \pm -\frac{Z_{99;0.995}}{\sqrt{100}}\cdot 6 = 27{,}5 \pm \frac{2.626}{10}\cdot 6 = 27{,}5 \pm 1{,}58,$$

а сам интервал (25,92; 29,08).

Замечаем, что интервал стал шире, что объясняется уменьшением объема имеющейся информации из-за незнания ещё одного параметра генеральной совокупности. ◆

§25. Методы получения оценок

До сих пор мы считали, что оценка неизвестного параметра известна и занимались изучением ее свойств с целью использования их при построении доверительного интервала. В этом параграфе рассмотрим вопрос о способах построения оценок.

1. Метод моментов

Пусть требуется оценить неизвестный параметр $\vec{\theta}$, вообще говоря, векторный, $\vec{\theta} = (\theta_1,...,\theta_k)$. При этом предполагается, что вид функции распределения известен с точностью до параметра $\vec{\theta}$,

$$F_\xi(x) = F(x,\vec{\theta}) = F(x,\theta_1,...,\theta_k).$$

В таком случае все моменты случайной величины ξ становятся функциями от $\vec{\theta}$:

$$\int_{-\infty}^{\infty} x^j dF(x,\theta_1,...,\theta_k) = \alpha_j(\theta_1,...,\theta_k).$$

Метод моментов требует выполнения следующих действий:

1. Вычисляем k «теоретических» моментов
$$\alpha_1(\theta_1,...,\theta_k),\ \alpha_2(\theta_1,...,\theta_k),...,\alpha_k(\theta_1,...,\theta_k)\ .$$
2. По выборке $x_1,...x_n$ строим k одноименных выборочных моментов. В излагаемом контексте это будут моменты
$$\mu_j(x_1,...,x_n) = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i^j,\ j = 1,2,...,k.$$
3. Приравнивая «теоретические» и одноименные им выборочные моменты, приходим к системе уравнений относительно компонент оцениваемого параметра
$$\begin{cases} \alpha_j(\theta_1,...,\theta_k) = \mu_j(x_1,...,x_n), \\ (j = \overline{1;k}). \end{cases} \qquad (25.1)$$
4. Решая полученную систему (точно или приближенно), находим исходные оценки $(\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k)$. Они, конечно, являются функциями от выборочных значений $x_1,...,x_n$.

Мы изложили порядок действий, исходя из начальных - теоретических и выборочных - моментов. Он сохраняется при ином выборе моментов, начальных, центральных или абсолютных, который определяется удобством решения системы (25.1) или ей подобной.

Перейдем к рассмотрению примеров.

Пример 25.1. Пусть случайная величина ξ распределена равномерно на отрезке $[\alpha;\beta]$, где α и β - неизвестные параметры. По выборке $(x_1,...,x_n)$ объема n из распределения случайной величины ξ. Требуется оценить α и β.

Р е ш е н и е. (примера 25.1)

В данном случае распределение определяется плотностью

$$f_\xi(x) = \begin{cases} 0, & x \notin (\alpha;\beta): \\ \dfrac{1}{\beta-\alpha}, & x \in [\alpha;\beta]. \end{cases}$$

1) Вычислим первые два начальных «теоретических» момента:

$$M\xi = \int_\alpha^\beta \frac{x dx}{\beta-\alpha} = \frac{\alpha+\beta}{2},$$

$$M\xi^2 = \int_\alpha^\beta \frac{x^2 dx}{\beta-\alpha} = \frac{1}{3}\cdot(\alpha^2 - \alpha\cdot\beta + \beta^2).$$

2) Вычислим по выборке два первых начальных выборочных момента

$$\mu_1 = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i,$$

$$\mu_2 = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i^2.$$

3) Составим систему уравнений

$$\begin{cases} \dfrac{1}{2}\cdot(\alpha+\beta) = \mu_1, \\ \dfrac{1}{3}\cdot(\alpha^2 - \alpha\cdot\beta + \beta^2) = \mu_2. \end{cases}$$

4) Из первого уравнения выразим α через β

$$\alpha = 2\mu_1 - \beta$$

и подставим во второе уравнение, в результате чего придём к квадратному уравнению

$$\beta^2 - 2\mu_1\beta + 2\mu_1^2 - 3\mu_2 = 0,$$

решая которое, находим два корня

$$\beta_1 = \mu_1 - \sqrt{3}\cdot\sqrt{\mu_2 - (\mu_1)^2}, \qquad \beta_2 = \mu_1 + \sqrt{3}\cdot\sqrt{\mu_2 - (\mu_1)^2}.$$

Соответствующие значения α таковы

$$\alpha_1 = \mu_1 + \sqrt{3}\cdot\sqrt{\mu_2 - (\mu_1)^2}, \qquad \alpha_2 = \mu_1 - \sqrt{3}\cdot\sqrt{\mu_2 - (\mu_1)^2}.$$

Поскольку по смыслу задачи должно выполнятся условие $\alpha < \beta$, выбираем в качестве решения системы и оценок неизвестных параметров

$$\hat{\alpha} = \mu_1 - \sqrt{3}\cdot\sqrt{\mu_2 - (\mu_1)^2}, \qquad \hat{\beta} = \mu_1 + \sqrt{3}\cdot\sqrt{\mu_2 - (\mu_1)^2}.$$

Замечая, что $\cdot\mu_2 - (\mu_1)^2$ есть не что иное, как выборочная дисперсия S^2, получаем окончательно

$$\hat{\alpha} = \bar{x} - \sqrt{3}\cdot S; \qquad \hat{\beta} = \bar{x} + \sqrt{3}\cdot S.$$

Если бы мы выбрали в качестве «теоретических» моментов математическое ожидание и дисперсию, $M\xi = \frac{\alpha+\beta}{2}$, $D\xi = \frac{(\beta-\alpha)^2}{12}$, то пришли бы к системе (с учетом неравенства $\alpha < \beta$)

$$\begin{cases} \dfrac{\alpha+\beta}{2} = \overline{x}, \\ \dfrac{\beta-\alpha}{2\sqrt{3}} = S, \end{cases}$$

которая линейна и решается проще предыдущей. Ответ, конечно, совпадает с уже полученным. ◆

Наконец, отметим, что наши системы всегда имеет решение и при том единственное. Полученные оценки, конечно, состоятельны, однако свойствам несмещенности не обладают.

2. Метод максимального правдоподобия

Изучается, как и прежде, случайная величина ξ, распределение которой задается либо вероятностями её значений $f(x,\vec{\theta})$, если ξ дискретна, либо плотностью распределения $f(x,\vec{\theta})$, если ξ непрерывна, где $\vec{\theta} = (\theta_1,...,\theta_k)$ - неизвестный векторный параметр. Пусть $(x_1,...x_n)$ - выборка значений ξ. Естественно в качестве оценки $\tilde{\vec{\theta}}$ взять то значение параметра, при котором вероятность получения уже имеющейся выборки максимальна.

Выражение

$$L(x_1,x_2,...,x_n,\vec{\theta}) = f(x_1,\vec{\theta})\cdot f(x_2,\vec{\theta})\cdot...\cdot f(x_n,\vec{\theta})$$

называют *функцией правдоподобия*, она представляет собой совместное распределение или совместную плотность случайного вектора с n независимыми координатами, каждая из которых имеет то же распределение (плотность), что и ξ.

В качестве оценки неизвестного параметра $\vec{\theta}$ берется такое его значение $\hat{\vec{\theta}}$, которое доставляет максимум функции $L(x_1,x_2,...,x_n,\vec{\theta})$, рассматриваемой как функции от $\vec{\theta}$ при фиксированных значениях $x_1,...x_n$. Оценку $\hat{\vec{\theta}}$ называют *оценкой максимального правдоподобия*. Заметим, что $\hat{\vec{\theta}}$ зависит от объема выборки n и выборочных значений $x_1,...x_n$

$$\hat{\vec{\theta}} = \hat{\vec{\theta}}(x_1,...x_n),$$

и, следовательно, сама является случайной величиной.

Отыскание точки максимума функции $L(x_1,x_2,...,x_n,\vec{\theta})$ представляет собой отдельную задачу, которая облегчается, если функция дифференцируема по параметру $\vec{\theta}$.

В этом случае удобно вместо функции $L(x_1,x_2,...,x_n,\vec{\theta})$ рассматривать её логарифм, поскольку точки экстремума функции и её логарифма совпадают.

Методы дифференциального исчисления позволяют найти точки, подозрительные на экстремум, а затем выяснить, в какой из них достигается максимум.

С этой целью рассматриваем вначале систему уравнений

$$\begin{cases} \dfrac{\partial \mathrm{Ln}\{L(x_1,\ldots,x_n,\theta_1,\ldots\theta_k)\}}{\partial\theta_i} = 0, \\ i = 1,2,\ldots,k\ , \end{cases} \qquad (25.2)$$

решения которой $(\hat{\theta}_1,\ldots,\hat{\theta}_k)$ - точки, подозрительные на экстремум. Затем по известной методике, вычисляя значения вторых производных

$$\begin{cases} \dfrac{\partial^2 \mathrm{Ln}\{L(x_1,\ldots,x_n,\theta_1,\ldots\theta_k)\}}{\partial\theta_i\partial\theta_j} = 0, \\ i,j = 1,2,\ldots,k\ , \end{cases}$$

по знаку определителя, составленного из этих значений, находим точку максимума.

Оценки, полученные по методу максимального правдоподобия, состоятельны, хотя могут оказаться смещенными.

Рассмотрим примеры.

Пример 25.2. Пусть производится некоторый случайный эксперимент, исходом которого может быть некоторое события A, вероятность P(A) которого неизвестна и подлежит оцениванию.

Р е ш е н и е.

Введем случайную величину ξ равенством

$$\xi = \begin{cases} 1, & \text{если событие A произошло,} \\ 0, & \text{если событие A не произошло (произошло событие } \overline{A}\text{).} \end{cases}$$

Распределение случайной величины ξ задается равенством

$$f(x,p) = p^x \cdot (1-p)^{1-x},\ \ x = 0,1.$$

Выборкой в данном случае будет конечная последовательность $(x_1,\ldots x_n)$, где каждое из x_i может быть равно 0 либо 1.

Функция правдоподобия будет иметь вид

$$L(x_1,\ldots,x_n,p) = p^{x_1}\cdot(1-p)^{1-x_1}\cdot\ldots\cdot p^{x_n}\cdot(1-p)^{1-x_n} = p^{\sum_{i=1}^{n} x_i}\cdot(1-p)^{n-\sum_{i=1}^{n} x_i}.$$

Найдем точку её максимума по p, для чего вычислим производную логарифма

$$\frac{d\ln L(x_1,\ldots,x_n,p)}{dp} = \frac{d}{dp}[\ln p\cdot\sum_{i=1}^{n} x_i + (n-\sum_{i=1}^{n} x_i)\cdot\ln(1-p)] = \frac{1}{p}\cdot\sum_{i=1}^{n} x_i - \frac{1}{1-p}\cdot(n-\sum_{i=1}^{n} x_i).$$

Обозначим $m = \sum_{i=1}^{n} x_i$ - это число равно количеству единиц «успехов» в выбранной последовательности.

Приравняем полученную производную к нулю

$$\frac{m}{p} - \frac{n-m}{1-p} = 0$$

и решим полученное уравнение

$$\hat{p} = \hat{p}(x_1,\ldots,x_n) = \frac{m}{n}.$$

Поскольку производная $\frac{m-np}{p(1-p)}$ меняет знак с «+» на «-» при возрастании p от 0 до 1, точка $\hat{p}$ есть точка максимума функции L, а $\hat{p}=\frac{m}{n}$ - оценка максимального правдоподобия параметра p. Заметим, что отношение $\frac{m}{n}$ есть частота появления события A в первых n испытаниях.

Поскольку m есть число «успехов» в последовательности n независимых испытаний (в схеме Бернулли), то $M\hat{p}=M\left(\frac{m}{n}\right)=\frac{np}{n}=p$, и $\hat{p}$ - несмещенная оценка. В силу закона больших чисел Бернулли $\frac{m}{n}$ стремится по вероятности к p, и оценка состоятельна. ◆

Пример 25.3. Построим оценки неизвестных математического ожидания и дисперсии нормально распределенной случайной величины ξ с параметрами a,σ^2.

Р е ш е н и е.

В условиях примера случайная величина определяется плотностью распределения

$$f(x;a,\sigma^2)=\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\ell^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Сразу выпишем логарифм функции правдоподобия

$$\ln L(x_1,\ldots,x_n;a,\sigma^2)=\ln\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n\cdot\ell^{-\frac{\sum_{k=1}^{n}(x_k-a)^2}{2\sigma^2}}=-\frac{n}{2}\cdot\ln(2\pi)-\frac{n}{2}\cdot\ln\sigma^2-\frac{\sum_{k=1}^{n}(x_k-a)^2}{2\sigma^2}.$$

Составим систему уравнений для нахождения экстремальных точек

$$\begin{cases}\dfrac{\partial\ln L}{\partial a}=\dfrac{\sum_{k=1}^{n}(x_k-a)}{\sigma^2}=0,\\ \dfrac{\partial\ln L}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\cdot\sum_{k=1}^{n}(x_k-a)^2=0.\end{cases}$$

Из первого уравнения находим $\hat{a}=\frac{1}{n}\cdot\sum_{k=1}^{n}x_k$, из второго, подставляя найденное значение $\hat{a}$, находим $\hat{\sigma}^2=\frac{1}{n}\cdot\sum_{k=1}^{n}(x_k-\hat{a})^2$.

Вычислим вторые производные функции lnL в точке $(\hat{a},\hat{\sigma}^2)$:

$$A=\frac{\partial^2\ln L}{\partial^2 a}=-\frac{n}{\hat{\sigma}^2},\qquad B=\frac{\partial^2\ln L}{\partial a\partial\sigma^2}=0,\qquad C=\frac{\partial^2\ln L}{\partial^2\sigma^2}=-\frac{n}{2\hat{\sigma}^2}.$$

Поскольку определитель

$$\Delta = AC - B^2 = \begin{vmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^2} \end{vmatrix} > 0,$$

а $A < 0$, то найденная точка в самом деле точка максимума функции правдоподобия.

Заметим, что оценка $\hat{a} = \frac{1}{n} \cdot \sum_{k=1}^{n} x_k$ есть выборочное среднее (несмещенная и состоятельная оценка математического ожидания), а $\hat{\sigma}^2 = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \hat{a})^2$ - выборочная дисперсия (смещенная оценка дисперсии). ◆

§26. Проверка статистических гипотез

Во многих практических ситуациях исследователю приходится принимать решение о том, какая из возможностей имеет место (продукция может быть бракованной или качественной, технологический процесс нарушается либо нет, точность обработки детали в пределах нормы, ниже нормы или выше её и т. п.). С общеметодологической точки зрения речь идет о выдвижении некоторой гипотезы, которая отвергается либо принимается после проведения некоторого эксперимента. Последний может иметь динамический либо статистический (стохастический) характер, и в этом случае говорят, что гипотеза является *статистической.*

Математическая постановка описанного выше процесса выглядит так.

Наблюдается случайная величина ξ, функция распределения которой $F(x,\theta)$ известна с точностью до параметра θ. При этом считаем, что θ принимает значение из некоторого множества Θ. Выделим в этом множестве некоторую подобласть, $\Theta_0 \subset \Theta$. Статистическая гипотеза состоит в том, что истинное значение параметра принадлежит Θ_0 . Решение о принятии либо отклонении гипотезы принимается на основании ряда наблюдений над значениями ξ, то есть на основании выборки $x_1, \ldots, x_n$.

В различных конкретных ситуациях эта задача решается по-разному, к обзору которых мы и перейдем.

• *Проверка простой гипотезы против простой альтернативы*

Выделим во множестве Θ всех возможных значений два одноэлементных подмножества $\Theta_0=\{\theta_0\}$ и $\Theta_1=\{\theta_1\}$. Относительно истинного параметра выдвинем две гипотезы: «нулевую», состоящую в том,

что $\theta = \theta_0$, и «альтернативу», состоящую в том, что $\theta = \theta_1$ соответственно. В этом частном случае гипотезы называют *простыми*.

Как уже говорилось, принятие или отклонение гипотезы **H_0** производится на основании статистической выборки $\bar{x} = (x_1, \ldots, x_n)$. Общее правило при этом выглядит следующим образом. Пусть $X = \{\bar{x}\}$ множество всех возможных выборок. Разобьем **X** на два непересекающихся подмножества **W** и **Q** : $X = W \cup Q$, $W \cap Q = \varnothing$. Если оказывается, что $\bar{x} \in W$, гипотеза H_0 отклоняется (по этой причине W называют *критической областью критерия*), если же $\bar{x} \in Q$, гипотеза принимается.

В силу того, что производимый эксперимент является стохастическим, выборочная точка $\bar{x}$ может оказаться в обеих областях, W и Q, независимо от того, верна ли нулевая гипотеза или альтернатива, что приводит к необходимости рассматривать различные возможности ошибочных решений.

Ошибкой первого рода называют решение об отклонении гипотезы H_0 в то время как она верна;

Ошибкой второго рода называют решение о принятии гипотезы H_0 в то время как она ошибочна.

С ошибками первого и второго рода связаны их вероятности:

обозначим α вероятность ошибки первого рода

$$\alpha = P_{\theta_0}\{\bar{x} \in W\},$$

и β - вероятность ошибки второго рода

$$\beta = P_{\theta_1}\{\bar{x} \in Q\},$$

где $P_\theta\{A\}$ означает вероятность события A, вычисленную по функции распределения $F(x,\theta)$.

Пусть случайная величина ξ обладает плотностью $f(x,\theta)$. Тогда совместная плотность выборки равна произведению плотностей, $f(\bar{x},\theta) = f(x_1,\ldots,x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$, и вероятности ошибок вычисляются по формулам

$$\alpha = \int_W \ldots \int f(x_1,\theta_0)\ldots f(x_n,\theta_0)dx_1\ldots dx_n$$

и

$$\beta = \int_Q \ldots \int f(x_1,\theta_1)\ldots f(x_n,\theta_1)dx_1\ldots dx_n$$

соответственно.

Если же ξ дискретна с распределением $f(x,\theta)$, то совместное распределение выборки имеет тот же вид, но интегралы в предыдущих формулах следует заменить кратными суммами.

Очевидно, что α и β на самом деле зависят от выбора критической области W и, желая уменьшить, например, вероятность ошибки первого рода

сужением W, мы в то же время увеличиваем вероятность ошибки второго рода.

Пример 26.1. Поставщик электроламп считает, что надежность его товара (вероятность того, что лампочка будет гореть в течение нормативного периода) равна 0,8. Потребитель же считает, что надежность равна 0,6.

В примере речь идет об испытании простой гипотезы H_0: $\theta = 0{,}6$ против альтернативы H_1: $\theta = 0{,}8$, где θ - надежность электролампы.

Р е ш е н и е.

С целью проверки нулевой гипотезы потребитель испытывает 10 лампочек и отвергает ее, если из них по меньшей мере 7 не выходят из строя в течение нормативного срока «жизни».

Пространство выборок состоит из $2^{10} = 1024$ строчек длинны 10, состоящей из символов 0 и 1 (0 означает, что лампа бракована, 1 - годна). Критическая область W состоит из тех строк, которые содержат не меньше 7 символов 1.

Вероятность ошибки первого рода (то есть ошибки, состоящей в том, что потребитель считает партию бракованной, в то время как она доброкачественна) равна

$$\alpha = \sum_{k=7}^{10} C_{10}^{k} \cdot 0.6^{k} \cdot 0.4^{10-k} \approx 0.38 \ .$$

Вероятность ошибки второго рода (состоит в том, что потребитель принимает бракованную партию) равна

$$\beta = \sum_{k=0}^{6} C_{10}^{k} \cdot 0.8^{k} \cdot 0.2^{10-k} \approx 0.13 \ .$$

Выберем критическую область несколько иначе: пусть W' состоит из тех строк, которые имеют в себе не менее 8 символов 1. Тогда вероятности ошибок примут иные значения:

$$\alpha' = \sum_{k=8}^{10} C_{10}^{k} \cdot 0.6^{k} \cdot 0.4^{10-k} \approx 0.16 \ ,$$

$$\beta' = \sum_{k=0}^{7} C_{10}^{k} \cdot 0.8^{k} \cdot 0.2^{10-k} \approx 0.33 . \quad \blacklozenge$$

Поскольку одновременно уменьшить вероятности ошибок первого и второго рода невозможно, поступают следующим образом.

Фиксируют ошибку первого рода α и подбирают из всех критических областей с данной α ту, для которой вероятность ошибки второго рода минимальна, или, что то же самое, максимальна *мощность* критерия, равная по определению $1 - \beta$.

В случае, когда существует плотность f(x,θ), для нахождения такого наиболее мощного критерия пользуются фундаментальной леммой Неймана - Пирсона.

Лемма. ***(Неймана-Пирсона)***. Критическая область критерия наибольшей мощности определяется так:

$$W^* = \left\{ \bar{x} = \frac{f(x_1,\theta_1)\ldots f(x_n,\theta_1)}{f(x_1,\theta_0)\ldots f(x_n,\theta_0)} \geq K_\alpha \right\},$$

где число K_α выбирается так, чтобы вероятность ошибки первого рода равнялась α.

Доказательство леммы мы приводить не будем, рассмотрим пример.

Пример 26.2. Детали изготовляются на двух равноточных станках, каждый из которых обладает своей систематической погрешностью. Отклонение размера изготовленной детали от эталонной подчиняется нормальному закону. На склад поступает партия деталей, изготовленных одним станком. Требуется выяснить, каким именно станком она выполнена.

Обозначим ξ отклонение размера детали от эталона. Условия примера означают, что случайная величина ξ распределена по нормальному закону, параметры которого таковы: математическое ожидание может принимать два значения θ_0 либо θ_1 (различные систематические погрешности), а дисперсия одна и та же (станки равноточные), примем ее за 1. Пусть для определенности $\theta_0 > \theta_1$

Речь идет о проверке гипотезы H_0: $\theta = \theta_0$ при альтернативе H_1: $\theta = \theta_1$ относительно параметра плотности

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}} \cdot \ell^{-\frac{(x-\theta)^2}{2}}.$$

Неравенство из леммы Неймана-Пирсона примет вид

$$\frac{(2\pi)^{-\frac{n}{2}} \cdot \prod_{i=1}^{n} \ell^{-\frac{(x-\theta_1)^2}{2}}}{(2\pi)^{-\frac{n}{2}} \cdot \prod_{i=1}^{n} \ell^{-\frac{(x-\theta_0)^2}{2}}} = \exp\left\{\frac{1}{2}\left[\sum_{i=1}^{n}(x_i-\theta_0)^2 - \sum_{i=1}^{n}(x_i-\theta_1)^2\right]\right\} =$$

$$= \exp\left\{\frac{1}{2}\left[2(\theta_1-\theta_2)\cdot\sum_{i=1}^{n}x_i + (\theta_0^2-\theta_1^2)\right]\right\} \geq K_\alpha.$$

Прологарифмировав обе части его, придем к виду:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq K'_\alpha ,$$

где $K'_\alpha = \theta_0 + \theta_1 + \frac{\ln K_\alpha}{n(\theta_1 - \theta_0)}$.

Поскольку проделанные преобразования эквивалентны, можно вместо отыскания K_α искать K'_α, не интересуясь конкретной связью между ними.

Зададим размер критерия α и вычислим K'_α исходя из равенства

$$\alpha = P\left\{\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq K'_\alpha\right\}.$$

Поскольку $\xi_1,...,\xi_n$ независимы и нормальны, то случайная величина $\frac{1}{n}\sum_{i=1}^{n}\xi_i$ (выборочное среднее) подчинена нормальному закону с параметрами $(\theta,\frac{1}{n})$, то есть

$$\alpha = \frac{\sqrt{1/n}}{\sqrt{2\pi}} \int_{K'_\alpha}^{\infty} \ell^{-\frac{(x-\theta_0)^2 n}{2}} dx .$$

Сделав замену $y = \sqrt{n}(x - \theta_0)$, приходим к равенству:

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{n}(K'_\alpha - \theta_0)}^{\infty} \ell^{-\frac{(y)^2}{2}} dy = \frac{1}{2} - \Phi(\sqrt{n}(K'_\alpha - \theta_0)) .$$

Если обозначить z_α корень уравнения $\Phi(z) = \frac{1}{2} - \alpha$, решаемого приближенно с помощью таблицы (прил. 2), получаем окончательно выражение для K'_α:

$$K'_\alpha = \theta_0 + \frac{z_\alpha}{\sqrt{n}} .$$

Итак, если выборочное среднее превышает K'_α, гипотезу H_0 следует отвергнуть (детали изготовлены, итак, вторым станком).

Оказывается, для построения наиболее мощного критерия вовсе не требуется задавать критическую область в пространстве X всех выборок - она приводится к полупрямой $[K'_\alpha,+\infty)$ на числовой оси, где откладываются значения выборочного среднего. ◆

Сделаем важное замечание относительно выбора размера критерия α.

На практике его выбирают так, чтобы событие, имеющее вероятность α, можно было считать «физически» невозможным, крайне редким.

- *Проверка гипотез о параметрах нормального закона*

В предыдущем разделе изучался вопрос о проверке простой гипотезы при простой альтернативе: множество параметров, соответствующее каждой из них, состоит из одной точки.

Гипотеза, которой соответствует множество параметров, состоящее более чем из одной точки, называется *сложной*. Общей теории проверки

гипотез, когда хотя бы одна из них сложная, не существует. Мы ограничимся рассмотрением некоторых частных случаев, связанных с гипотезами о параметрах нормального закона.

Пример 26.3. Гипотеза о значении математического ожидания нормального закона при известной дисперсии.

Проверяется простая гипотеза Но: $\theta = \theta_0$ о параметре нормального закона с плотностью

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot \ell^{-\frac{(x-\theta)^2}{2\sigma^2}}$$

против сложной альтернативы H_1: $\theta \neq \theta_0$.

Известно из предыдущего, что выборочное среднее в выборке из нормального закона с параметрами (a,σ^2) имеет нормальное распределение с параметрами $(a,\frac{\sigma^2}{n})$, поэтому в рассматриваемом примере нормированное выборочное среднее $\frac{(\bar{x}-\theta_0)\sqrt{n}}{\sigma}$ при условии, что гипотеза H_0 верна, будет иметь нормальное распределение с параметрами (0,1). ◆

Используя этот факт, можно построить критерий для проверки гипотезы следующим образом.

1. Выберем размер критерия α;
2. Вычислим число $Z_{\frac{1-\alpha}{2}}$, исходя из равенства

$$P\left\{\left|\frac{(\bar{x}-\theta_0)\sqrt{n}}{\sigma}\right| > Z_{\frac{1-\alpha}{2}}\right\} = \frac{2}{\sqrt{2\pi}}\cdot\int_{Z_{\frac{1-\alpha}{2}}}^{\infty}\ell^{-\frac{x^2}{2}}dx = \alpha,$$

то есть $Z_{\frac{1-\alpha}{2}}$ есть корень уравнения $\Phi(Z_{\frac{1-\alpha}{2}}) = \frac{1-\alpha}{2}$, решаемого с помощью таблицы значений $\Phi(z)$ (прил. 2);

3. Вычислим по выборке $\bar{x}$, а затем статистику $z = \frac{(\bar{x}-\theta_0)\sqrt{n}}{\sigma}$;
4. Если выполняется равенство $|Z| > Z_{\frac{1-\alpha}{2}}$, то в силу принципа физической невозможности, гипотезу H_0 следует отвергнуть.

Численный пример. Точность обработки детали характеризуется среднеквадратическим отклонением (взятым из государственного стандарта) $\sigma = 1$. Отклонение размера детали от нормативного, равного 20, подчиняется нормальному распределению. В результате измерения 25 деталей, изготовленных станком, получено выборочное среднее $\bar{x} = 19{,}85$. Требуется проверить на уровне значимости $\alpha = 0{,}01$ гипотезу о том, что средний размер детали соответствует нормативному. По таблице (прил.2) находим

$Z_{\frac{1-\alpha}{2}} = Z_{0.495} = 2{,}576$. Поскольку $\left|\frac{(\bar{x}-\theta_0)\sqrt{n}}{\sigma}\right| = 1{,}25$ и $1{,}25<2{,}576$, то нулевая гипотеза принимается. ◆

Пример 26.4. Проверка гипотезы о равенстве средних двух независимых нормально распределенных случайных величин.

Пусть имеем две независимые нормальные случайные величины ξ и η, распределенные с параметрами (θ_1,σ^2) и (θ_2,σ^2), причем σ будем считать известным.

Требуется проверить сложную нулевую гипотезу H_0: $\theta_1 = \theta_2$ при сложной альтернативе H_1: $\theta_1 \neq \theta_2$.

Обозначим $(x_1,x_2,...,x_n)$, $(y_1,y_2,...,y_n)$ выборки из законов ξ, η соответственно.

В предположениях примера случайная величина $\bar{x}-\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{m}\sum_{j=1}^{m} y_j$ подчиняется нормальному закону с параметрами $(\theta_1-\theta_2, \frac{\sigma^2}{n}+\frac{\sigma^2}{m})$, а нормированная разность $z = \frac{[\bar{x}-\bar{y}-(\theta_1-\theta_2)]\cdot\sqrt{\frac{mn}{m+n}}}{\sigma}$ имеет нормальное распределение с параметрами (0,1).

В предположении, что гипотеза H_0 справедлива, статистика z принимает вид

$$z = \frac{(\bar{x}-\bar{y})}{\sigma}\cdot\sqrt{\frac{mn}{m+n}}.$$

Дальнейшие рассуждения аналогичны рассуждениям примера 26.3.

1. Выбираем размер критерия α.
2. Выбираем число $Z_{\frac{1-\alpha}{2}}$.
3. Вычисляем по выборке $\bar{x}$, $\bar{y}$, z.
4. Если $|Z| > Z_{\frac{1-\alpha}{2}}$, то H_0 отклоняется.

Численный пример. Сравниваем два теодолита, каждый из которых имеет заданное стандартом среднеквадратическое отклонение 0,3°. Отклонения в измерениях угла от истинного его значения подчиняются нормальному закону. Произведено соответственно 25 и 50 измерений первым и вторым теодолитом, вычислены соответственно выборочные средние $\bar{x}$ = = 9,79°, $\bar{y}$ = 9,60°. На уровне значимости α = 0,01 выяснить, отличаются ли теодолиты.

Как и в примере 12.2.1 $Z_{\frac{1-\alpha}{2}} = Z_{0.495} = 2{,}576$.

Значение статистики

$$|Z| = \frac{\left|\bar{x} - \bar{y}\right|}{\sqrt{\frac{\sigma\xi^2}{n_1} + \frac{\sigma\eta^2}{n_2}}} = \frac{9{,}79 - 9{,}60}{0{,}3 \cdot \sqrt{\frac{1}{25} + \frac{1}{50}}} = \frac{0{,}19}{0{,}3} \cdot \sqrt{\frac{1250}{75}} \cong 2{,}586.$$

Поскольку 2,586 > 2.576 , гипотеза о том, что теодолиты не отличаются, должна быть отвергнута. ◆

З а м е ч а н и е. Однако следует отметить, что для значений $|Z| \le 2{,}576$, еще нельзя утверждать, что гипотеза подтвердилась:
можно только признать допустимость гипотезы для рассмотренных выборочных наблюдений до тех пор, пока более обстоятельные исследования не позволят сделать противоположное заключение.

Следовательно, с помощью проверки статистических гипотез можно лишь отвергнуть проверяемую гипотезу, но никогда нельзя доказать ее справедливость.

§27. Критерий согласия χ^2 Пирсона

(*проверка гипотез о законе распределения*)

До сих пор мы предполагали, что вид функции распределения случайной величины ξ известен. Такое знание может быть получено из предыдущего опыта или по результатам аналогичных исследований других авторов.

Однако в любом случае такое знание нельзя считать точным, а лишь исходным материалом для последующего изучения с учетом конкретных обстоятельств явления, находящегося в рассмотрении.

Задача ставится следующим образом.

Изучается случайная величина ξ, относительно которой выдвигается гипотеза о том, что ее функция распределения есть F(x).

Критерии, предназначенные для проверки такой гипотезы, называются *критериями согласия*.

Мы изложим один из них, называемый *критерием* χ^2 (хи-квадрат) или *критерием К. Пирсона* по имени его автора.

Разобьем множество всех возможных значений случайной величины ξ точками $y_1 < y_2 < y_3 < ... < y_{r-1}$, в результате получим **r** интервалов $(-\infty, y_1], (y_1, y_2], ... (y_{r-1}, +\infty)$.

Имея статистическую выборку $x_1, x_2, ..., x_n$, вычислим числа $n_1, n_2, ..., n_r$, представляющие собой количества выборочных наблюдений, попавших в первый, второй, ..., r-й интервал. Разумеется, $\sum_{i=1}^{r} n_i = n$.

Считая, что поверяемая гипотеза верна, вычислим вероятности попадания случайной величины ξ в указанные выше интервалы (функцию распределения считаем непрерывной):

$$P_1 = F(x_1), P_2 = F(x_2) - F(x_1), \ldots, P_r = 1 - F(x_{r-1}).$$

В силу закона больших чисел следует ожидать, если гипотеза верна, что частоты n_i/n близки к P_i, на этой идее и основан критерий.

Рассмотрим статистику

$$\chi^2 = \sum_{i=1}^{r} \left(\frac{n_i - n \cdot p_i}{\sqrt{n \cdot p_i}} \right)^2 .$$

В курсах математической статистики доказывается, что статистика χ^2 при $n \to \infty$ стремится к случайной величине, распределенной по закону χ^2_r с r-1 степенями свободы, плотность которого имеет вид

$$S_r(x) = \frac{1}{2^{\frac{r}{2}} \cdot \Gamma(\frac{r}{2})} \cdot x^{\frac{r}{2}-1} \cdot \ell^{-\frac{x}{2}}, \quad x > 0.$$

Значения интеграла $\int_x^\infty S_r(y)dy$ табулированы для различных значений r (прил. 4).

Также доказывается, что если по выборке оценено k параметров для уточнения вида F(x), то предельным будет распределение χ^2_{r-k-1}, где l = r - k-1 число *степеней свободы*.

Критерий согласия строится следующим образом.

1. Выбираем уровень значимости α;
2. По таблице прил. 4 по данным r (и k, если по выборке оценивается несколько параметров) и α находим число $\chi^2_{r-1,\alpha}$ ($\chi^2_{r-k-1,\alpha}$) такое, что

$$\alpha = \int_{\chi^2_{r-1,\alpha}}^{\infty} S_{r-1}(y)dy ,$$

или

$$\alpha = \int_{\chi^2_{r--k1,\alpha}}^{\infty} S_{r-k-1}(y)dy .$$

3. По выборке находим числа n_i, вычисляем P_i и статистику χ^2;
4. Если выполнено $\chi^2 > \chi^2_{r-1,\alpha}$ ($\chi^2 > \chi^2_{r-k-1,\alpha}$), то гипотеза о согласии выборочных наблюдений с законом F(x) отвергается.

Пример 27.1. Для выяснения того, хорошо ли отбалансирована стрелка компаса, 500 раз произведено следующее исследование: по концу стрелки производится удар и после ее остановки измеряется угол

между начальным и конечным положением, отсчитываемый против движения часовой стрелки.

Все возможные значения угла (от 0° до 360°) разбиты на 12 равных интервалов, выборочные данные сгруппированы, в результате чего получена таблица

Таблица 27.1

Интервалы	**Градусы**											
	0-30	30-60	60-90	90-120	120-150	150-180	180-210	210-240	240-270	270-300	300-330	330-360
ni	41	34	54	39	49	45	41	33	37	41	47	39

Требуется проверить на уровне значимости α =0,1 гипотезу о том, что угол между начальным и конечным положением стрелки распределен равномерно на отрезке [0° , 360°].

Согласно гипотезе вероятности P_i все равны между собой,

$$P_i = \frac{1}{12}, (i = 1,2,...,12).$$

По таблице прил. 4 для 11 степеней свободы и $\alpha = 0,1$ находим

$$\chi^2_{11;0,1} = 17,3.$$

Вычисляем по выборочным данным значение статистики χ^2, которое оказывается равным 9,9966. Поскольку 9,9966<17,3, гипотеза о равномерности распределения принимается. ◆

§28. Основы корреляционного и регрессионного анализа

Целью любого исследования, осуществляемого в настоящее время, является использование его результатов в будущем, или, иначе говоря, прогнозирование состояния изучаемого явления. Примерами такого прогнозирования заполнены учебники всех естественнонаучных и экономических дисциплин. При этом, желая изучать явление во взаимосвязи с другими явлениями или величинами, приходится выделять некоторые из них, влияющие на изучаемое, оценивать степень и «качество» влияния, то есть характер связи между изучаемым (основным в данном исследовании) и влияющими на него величинами качественного или количественного характера.

В дальнейшем мы «основную», изучаемую, величину будем называть *зависимой переменной* и обозначать литерой y, прочие, влияющие на y,

величины будем называть *независимыми переменными* и обозначать литерами $x_1, x_2, ..., x_k$. Как y, так и $x_1, x_2, ..., x_k$, будем считать числовыми.

Различают два вида связей.

Если значение зависимой переменной становится известным, как только известны значения независимых переменных, говорят о связи *динамической* или *функциональной*, поскольку в этом случае существует закон, по которому вычисляется y в зависимости от $x_1, x_2, ..., x_k$, $y = f(x_1, x_2, ..., x_k)$. Примеры таких связей: закон свободного падения тела; закон Ома; закон Бойля-Мариотта; связь между стоимостью единицы товара и ценой, уплаченной за партию его; зависимость производительности труда и затрат рабочего времени.

Иначе обстоит дело, когда по значениям независимых величин можно установить лишь некоторую «среднюю» тенденцию в значениях зависимой переменной. Так, например, общепонятно, что между ростом человека и его весом существует зависимость, созданы таблицы такой зависимости, учитывающие еще и пол, и возраст, однако пользоваться ими можно лишь, опять же, «в среднем». Подобного рода связи называют *корреляционными* (от слова correlatio - соотношение - латынью), а задачей установления математической формы корреляционной связи занимается *регрессионный анализ*. Зависимая переменная y при этом рассматривается как случайная величина, а независимые переменные можно прямо или косвенно контролировать. *Корреляционный анализ* изучает совместное распределение всех измеряемых переменных с анализом точности оценивания одних величин через другие.

В отличие от функциональной связи в регрессионном анализе речь идет об установлении функции регрессии $M(y/x_1, x_2, ..., x_k) = f(x_1, ..., x_k)$, где символ $M(\cdot/\cdot)$ обозначает математическое ожидание случайной величины y при заданных значениях независимых переменных.

Здесь важно заметить следующее.

В то время как независимые переменные $x_1, x_2, ..., x_k$ контролируемы, управляемы, а y является случайной величиной, то по данным эксперимента, в котором $x_1, x_2, ..., x_k$ приняли вполне конкретные значения, можно судить лишь об оценке параметра, связанного с распределением y, оценок же, как мы уже знаем, можно построить много.

С точки зрения дальнейших применений желательно иметь оценку как можно более простого вида и которая удовлетворяла бы некоторому критерию оптимальности (подобному несмещенности, например, для оценок параметров).

Из всех элементарных функций (исключая константу) наиболее простой является линейная, этот случай мы и изучим в дальнейшем детально как наиболее прозрачный с точки зрения идейной и в то же время дающий возможность для дальнейших обобщений.

§29. Линейная регрессия и метод наименьших квадратов

Опишем вначале математическую постановку задачи, считая, что изучается одна зависимая переменная y в присутствии одной независимой переменной x (так называемая задача *парной регрессии*).

Пусть зависимость между x и y имеет вид

$$y = a_0 + a_1 x + \varepsilon,$$

где a_0, a_1 - постоянные коэффициенты, называемые *параметрами модели*, ε-случайная величина с математическим ожиданием 0 и дисперсией σ^2.

В этом случае уравнение регрессии превращается в уравнение прямой

$$\bar{y}(x) = M(y/x) = a_0 + a_1 \cdot x.$$

Предположим, что независимой переменной придали значения $x_1, x_2, \ldots, x_n$, в результате чего зависимая переменная приняла значения $y_1, y_2, \ldots, y_n$. В предположении линейной зависимости получаем n равенств

$$y_i = a_0 + a_1 x_i + \varepsilon_i, \quad i = \overline{1, n},$$

где ε_i - независимы и распределены так же, как ε.

Требуется по значениям пар (x_i, y_i) оценить неизвестные a_0, a_1.

Как мы уже знаем, каждая задача оценивания связана с некоторым критерием качества. В излагаемой нами теории таким критерием является критерий наименьших квадратов: $\sum_{i=1}^{n} \varepsilon_i^2 - \min$.

Запишем эту сумму иначе, так, чтобы была видна зависимость от a_0, a_1:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [\bar{y}(x_i) - y_i]^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2.$$

Теперь окончательно приходим к следующей задаче:

отыскать такие значения неизвестных параметров a_0, a_1, чтобы функция

$$Q(a_0, a_1) = \sum_{i=1}^{n} \left[y_i - a_0 - a_1 \cdot x_i \right]^2$$

приняла наименьшее значение.

Метод решения этой задачи известен из курса высшей математики.

Находим частные производные функции Q и приравниваем их к нулю, в результате чего приходим к системе линейных уравнений

$$\begin{cases} \dfrac{\partial Q}{\partial a_0} = -2 \cdot \sum\limits_{i=1}^{n} (y_i - a_0 - a_1 \cdot x_i) = 0, \\ \dfrac{\partial Q}{\partial a_1} = -2 \cdot \sum\limits_{i=1}^{n} (y_i - a_0 - a_1 \cdot x_i) \cdot x_i = 0. \end{cases}$$

После очевидных преобразований получаем систему

$$\begin{cases} n \cdot a_0 + a_1 \cdot \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ a_0 \cdot \sum\limits_{i=1}^{n} x_i + a_1 \cdot \sum\limits_{i=1}^{n} x_i^{\,2} = \sum\limits_{i=1}^{n} x_i \cdot y_i. \end{cases}$$

Обозначим выборочные средние

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i, \quad \overline{x^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} x^2{}_i.$$

В этих обозначениях после деления каждого уравнения системы на n она примет вид

$$\begin{cases} a_0 + a_1 \cdot \bar{x} = \bar{y}, \\ a_0 \cdot \bar{x} + a_1 \overline{x^2} = \overline{xy}, \end{cases}$$

а ее решение (искомые оценки коэффициентов уравнения регрессии) будет таким

$$\hat{a}_0 = \frac{\overline{x^2} \cdot \bar{y} - \overline{xy} \cdot \bar{x}}{\overline{x^2} - (\bar{x})^2},$$

$$\hat{a}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}.$$

Если ввести еще обозначение $S_x^2 = \overline{x^2} - \left(\overline{x^2}\right)$ и преобразовать выражение для $\hat{a}_0$:

$$\hat{a}_0 = \frac{\overline{x^2} \cdot \bar{y} - \overline{xy} \cdot \bar{x} \pm \bar{y} \cdot (\bar{x})^2}{S_x^2} = \frac{\bar{y} \cdot S_x^2 - \bar{x}(\overline{xy} - \bar{x} \cdot \bar{y})}{S_x^2} = \bar{y} - \hat{a}_1 \cdot \bar{x},$$

то оценка функции регрессии примет вид

$$\hat{\bar{y}}(x) = \hat{a}_0 + \hat{a}_1 x = \bar{y} - \hat{a}_1 \cdot \bar{x} + \hat{a} \cdot x = \bar{y} + \hat{a}_1 (x - \bar{x}).$$

Пример 29.1. Агент по продаже домов изучает зависимость между ценой дома **y** (в $ 1000) и общей его площадью **x** (в сотнях квадратных футов). С этой целью он произвел выборку из 15 домов и зафиксировал такие результаты:

Таблица 29.1

i	$\mathbf{x_i}$	$\mathbf{y_i}$	**i**	$\mathbf{x_i}$	$\mathbf{y_i}$
1	20.0	89.5	**9**	24.3	119.9

2	14.8	79.9	**10**	20.2	87.6
3	20.5	83.1	**11**	22.0	112.6
4	12.5	56.9	**12**	19.0	120.8
5	18.0	66.6	**13**	12.3	78.5
6	14.3	82.5	**14**	14.0	74.3
7	27.5	126.3	**15**	16.7	74.8
8	16.5	79.3			

Нанеся пары (x_i, y_i) на координатную плоскость, он получает так называемое *корреляционное облако*, вид которого позволяет предположить, что линейная зависимость между переменными не лишена оснований.

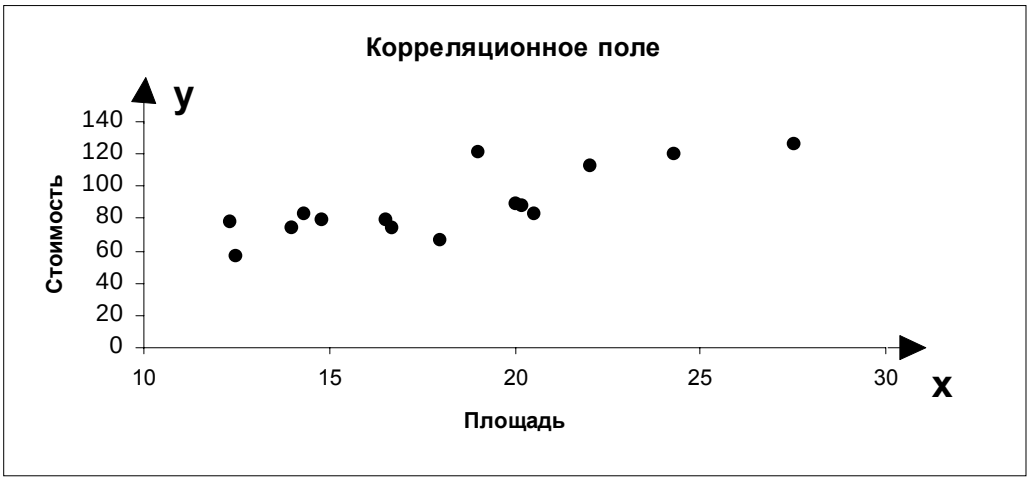


Рис. 29.1

Приняв эту гипотезу, он вычисляет

$$\bar{x} = 18.173; \quad \bar{y} = 88.8; \quad \overline{xy} = 1683.86; \quad \overline{x^2} = 348.15; \quad S_x^2 = 17.88;$$

а затем по полученным выше формулам оценки

$$\hat{a}_1 = 3.88; \quad \hat{a}_0 = \bar{y} - \hat{a}_1 \cdot \bar{x} = 18.354 .$$

Теперь прямая регрессии имеет уравнение

$$\hat{\bar{y}}(x) = 18{,}354 + 3{,}88x .$$

Ее график нанесем на корреляционное поле (рис. 29.2)

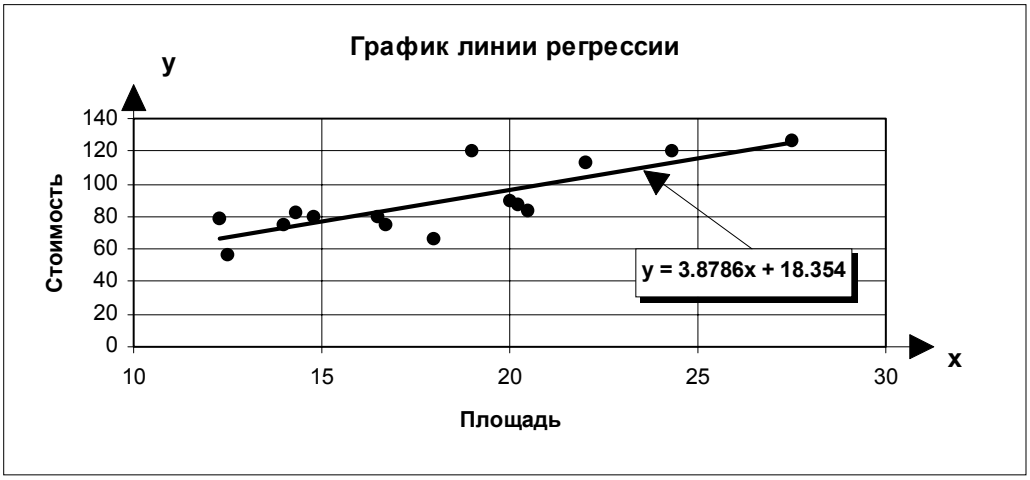


Рис. 29.2

◆

§30. Анализ коэффициентов уравнения регрессии при известном σ^2

Можно показать, что оценки $\hat{a}_0$ и $\hat{a}_1$ коэффициентов a_0 и a_1 являются несмещенными независимо от того, как распределены случайные «добавки» $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$.

Для получения более точных сведений о свойствах оценок предположим, что $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ распределены нормально с указанными ранее параметрами, причем следует различать случаи, когда σ^2 известно или нет.

Если σ^2 известно, то доказывается, что оценки распределены нормально, а их дисперсии равны

$$D(\hat{a}_0) = \frac{\overline{x^2}}{nS_x^2} \cdot \sigma^2$$

и

$$D(\hat{a}_1) = \frac{\sigma^2}{nS_x^2}.$$

Обладая указанными сведениями, можно строить доверительные интервалы для a_0 и a_1, а также производить проверку гипотез относительно их значений.

Пример 30.1 (*продолжение примера 29.1*). Допустим, что σ^2 известно и равно 169.

В этом предположении имеем

$$D(\hat{a}_0) = \frac{348{,}15}{15 \cdot 17{,}88} \cdot 169 = 219{,}38 \ ,$$

$$D(\hat{a}_1) = \frac{169}{15 \cdot 17{,}88} = 0{,}63 \ ,$$

а 95%-е доверительные интервалы будут таковы:

$$\hat{a}_0 - z_{0.475} \cdot \sqrt{219.38} < a_0 < \hat{a}_0 + z_{0.475} \cdot \sqrt{219.38}$$

и

$$\hat{a}_1 - z_{0.475} \cdot \sqrt{0{,}63} < a_1 < \hat{a}_1 + z_{0.475} \cdot \sqrt{0{,}63} \ ,$$

где $z_{0.475} = 1{,}96$ (см. таблицу прил. 1).

То есть

$$-10{,}69 < a_0 < 47{,}37$$

и

$$2{,}54 < a_1 < 5{,}22 \ .$$

Проверим гипотезу H_0: $a_1=0$ (она означает, что между x и y нет линейной связи) против альтернативы H_1 : $a_1 \neq 0$, при размере критерия $\alpha=0{,}05$.

В этом случае критическая область представляет собой внешность интервала $(-z_{0.475} \cdot \sqrt{D(\hat{a}_1)};\ +z_{0.475} \cdot \sqrt{D(\hat{a}_1)} = (-1{,}56,\ 1{,}56))$. Поскольку экспериментальное значение $\hat{a}_1 = 3{,}88$ выходит за его границу, нулевая (H_0) гипотеза отвергается. ◆

Сделаем еще одно важное замечание относительно дисперсии $\hat{a}_1$.

Желая сделать оценку коэффициента a_1 (называемого коэффициентом регрессии*) как можно точней, следует сделать ее дисперсию как можно меньше. Последнее ввиду равенства*

$$S_x^2 = \overline{x^2} - \left(\overline{x^2}\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

означает, что значения независимой переменной следует выбирать на границах интервала ее изменения. То есть, если хотим произвести 4 наблюдения, то два из них следует произвести при $x=x^{\backslash}$, и два при $x=x^{\backslash\backslash}$, где $[x^{\backslash}, x^{\backslash\backslash}]$ - отрезок допустимых значений контролируемой переменной.

§31. Оценивание σ^2

Полученная по методу наименьших квадратов оценка линии прямой регрессии является наилучшей, однако это вовсе не означает, что в действительности связь между x и y линейна. Судить о качестве оценивания можно по величине

$$RSS = \sum_{i=1}^{n} (y_i - \hat{a}_0 - \hat{a}_1 x_i)^2 \ ,$$

представляющей собой наименьшее значение функции Q. Обозначение RSS является стандартным в компьютерных программах (например, в пакете Microsoft Excel - SSresid) и образовано начальными литерами выражения *residual sum squares* - остаточная сумма квадратов (разностей между настоящими значениями y_i и значениями зависимой переменной, если бы регрессия была точной).

Доказывается, что частное $\frac{RSS}{n-2}$ является несмещенной оценкой σ^2.

Найдем выражение для RSS, удобное для вычислений:

$$RSS = \sum_{i=1}^{n}(y_i - \bar{y} + \hat{a}_1\bar{x} - \hat{a}_1 x_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2\cdot\hat{a}_1\sum_{i=1}^{n}(y_i - \bar{y})(\bar{x} - x_i) + \hat{a}_1^2\cdot\sum_{i=1}^{n}(\bar{x} - x_i)^2 =$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\cdot\hat{a}_1(\overline{xy} - \bar{x}\cdot\bar{y}) + \hat{a}_1^2\sum_{i=1}^{n}(\bar{x} - x_i) = n\cdot S_y^2 - \frac{(\overline{xy} - \bar{x}\cdot\bar{y})^2\cdot n}{S_x^2} =$$

$$= n\cdot\left[S_y^2 - \frac{(\overline{xy} - \bar{x}\cdot\bar{y})^2}{S_x^2}\right],$$

где $S_y^2 = \overline{y^2} - (\bar{y})^2$.

Пример 31.1 (*продолжение примера 29.1*). Считая теперь неизвестным, σ^2, вычислим его оценку.

Имеем по данным примера $\overline{y^2} = 8307{,}89$; $S_y^2 = 415{,}35$; $RSS = 2195{,}88$. Наконец, обозначив S^2 оценку для σ^2 получаем:

$$S^2 = \frac{RSS}{n-2} = \frac{2195{,}88}{13} = 168{,}91. \quad \blacklozenge$$

Если считать модель $y = a_0 + a_1 x + \varepsilon$ верной, то $D[y(x)] = \sigma^2$, и оценка для σ^2, построенная выше, дает возможность судить о «качестве» модели, сравнивая S^2 со средним $\bar{y}$.

§32. Анализ коэффициентов уравнения регрессии при неизвестном σ^2

При неизвестном σ^2 дисперсии оценок $\hat{a}_0$ и $\hat{a}_1$ заменяются их оценками:

- оценка дисперсии $\hat{a}_0 = \frac{\overline{x^2}}{nS_x^2}\cdot S^2$,
- оценка дисперсии $\hat{a}_1 = \frac{S^2}{nS_x^2}$.

Указанные оценки дисперсий можно использовать для построения доверительных интервалов и проверки гипотез относительно параметров модели, следует лишь при этом опираться не на нормальное распределение, а на распределение Стьюдента с числом степеней свободы n-2.

Так, если $\alpha \approx 0$, то доверительные интервалы будут иметь вид

- для a_o:

$$\hat{a}_0 \pm t(n-2,\ 1-\tfrac{1}{2}\alpha)\cdot\left(\frac{\overline{x^2}}{nS_x^2}\right)^{\frac{1}{2}}\cdot S,$$

- для a_1:

$$\hat{a}_1 \pm t(n-2,\ 1-\tfrac{1}{2}\alpha)\cdot\left(\frac{S^2}{nS_x^2}\right)^{\frac{1}{2}},$$

где $t(n-2,\ 1-\frac{1}{2}\alpha)$ - $(1-\frac{1}{2}\alpha)$- процентная точка распределения Стьюдента с числом степей свободы n-2.

Пример 32.1 (*продолжение примера 29.1*). Построим доверительные интервалы уровня доверия 0,95 для параметров a_0 и a_1, считая σ^2 неизвестным и заменив его оценкой S^2=168,91.

В этом случае t(13; 0.975) = 2,16 и доверительный интервал для a_0 будет таким:

- (-13,67; 50,35),

а для a_1 :

- (2,17; 5,59).

Как видим, оба интервала расширились, что объясняется уменьшением объема информации об условиях эксперимента. ◆

Проверка гипотезы H_0: $a_1 = 0$ против альтернативы H_1 : $a_1 \neq 0$ основывается на статистике

$$t = \frac{\hat{a}_1}{S\cdot(n\cdot S_x^2)^{-\frac{1}{2}}},$$

при этом критическая область имеет вид

$$|t| > t(n-2,\ 1-\tfrac{1}{2}\alpha)\ .$$

Пример 32.2 (*продолжение примера 29.1*). Проверим гипотезу H_0: $a_1=0$ против альтернативы H_1 : $a_1\neq 0$, при размере критерия α=0,05 в нашей задаче об агенте по продаже недвижимости.

Все необходимые вычисления уже нами сделаны, остается лишь найти значение статистики **t** :

- $t = \dfrac{3{,}88}{0{,}79} = 4{,}91.$

Поскольку 4,91 больше 2,16 , нулевая гипотеза отвергается. ◆

§33. Применение уравнения регрессии

Предсказание значения y *при данно* x

Уравнение регрессии может быть использовано с двух точек зрения:

- как отражение уже наблюдавшегося явления и
- как способ предсказания его будущего.

Ниже мы остановимся на втором аспекте.

Желая предсказать индивидуальное значение y при данном значении x, следует исходить из того, что оценка среднеквадратического отклонения y при данном x имеет вид

$$S\cdot\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{n\cdot S_x^2}},$$

что приводит к интервалу предсказания уровня α вида

$$\hat{y}\pm t(n-2,\ 1-\frac{\alpha}{2})\cdot S\cdot\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{n\cdot S_x^2}},$$

где $\hat{y}=\hat{a}_0+\hat{a}_1\cdot x$.

Если же нас интересует предсказание не частного значения y, а всего лишь среднего его значения (имея в виду, что речь идет об условном среднем y при данном x), то соответствующая оценка среднеквадратического отклонения M(y/x) имеет вид

$$S\cdot\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{n\cdot S_x^2}},$$

а соответствующий доверительный интервал уровня α будет таким:

$$\hat{y}\pm t(n-2,\ 1-\frac{\alpha}{2})\cdot S\cdot\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{n\cdot S_x^2}}.$$

Как видим, второй доверительный интервал уже первого, что и понятно: в первом случае речь идет о частном значении признака y, а во втором - лишь о его среднем значении.

Пример 33.1 (*окончание примера 29.1*). Пусть агент желает предсказать цену дома общей площадью 20 сотен квадратных футов.

Используя полученное уравнение регрессии

$$\hat{\bar{y}}(x)=18{,}354+3{,}88x,$$

он находит, что $\hat{y}=18{,}354+3{,}88\cdot 20=95{,}954$ (тыс. долл.). Однако, это всего лишь оценка, которая без указания на возможные колебания цены мало о чем говорит.

Найдем интервал предсказания уровня 0,95:

$$95{,}954\pm 2{,}16\cdot\sqrt{168{,}91}\cdot\sqrt{1+\frac{1}{15}+\frac{(20-18{,}17)^2}{268{,}19}}$$

или (66,792; 125,116).

Если агента интересует предсказание среднего значения цены большой совокупности домов с общей площадью 2000 квадратных футов, он воспользуется доверительным интервалом (с тем же уровнем доверия):

$$95{,}954 \pm 2{,}16 \cdot \sqrt{168{,}91} \cdot \sqrt{\frac{1}{15} + \frac{(20 - 18{,}17)^2}{268{,}19}}$$

или (88,056; 103,852).

Все сказанное выше хорошо видно на такой диаграмме (см. рис. 33.1)

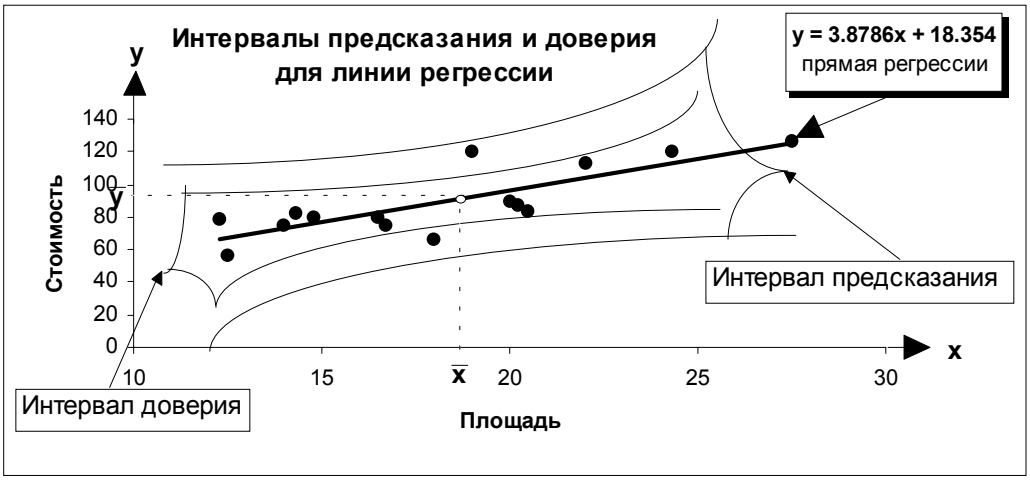


Рис. 33.1

◆

§34. Коэффициент корреляции

До сих пор мы занимались вопросом установления формы линейной зависимости между двумя признаками. Построенный критерий проверки гипотезы о равенстве нулю коэффициента регрессии a_1 дает возможность принять либо отвергнуть ее. В случае отклонения мы все же не можем сказать, насколько хороша выбранная линейная модель и оправдана ли она вообще. В настоящем параграфе мы приведем одну характеристику линейной связи между двумя случайными величинами.

Рассмотрим случайный вектор (ξ,η) и вычислим $M[(\eta - M\eta) + a(\xi - M\xi)]^2$, где **a** - числовой параметр:

$$f(a) = M[(\eta - M\eta) + a(\xi - M\xi)]^2 = M(\eta - M\eta)^2 + 2a \cdot M[(\eta - M\eta)(\xi - M\xi)] +$$
$$+ a^2 \cdot M(\xi - M\xi)^2 = D\eta + 2a \cdot M[(\eta - M\eta)(\xi - M\xi)] + a^2 \cdot D\xi$$

Видим, что f(a) есть квадратный трехчлен относительно **a**, принимающий только неотрицательные значения, так что его дискриминант неположителен, то есть

$$[M(\eta - M\eta)(\xi - M\xi)]^2 - D\eta \cdot D\xi \le 0,$$

или

$$\frac{[M(\eta - M\eta)(\xi - M\xi)]^2}{D\eta \cdot D\xi} \le 1,$$

или, наконец,

$$\frac{|M(\eta - M\eta)(\xi - M\xi)|}{\sqrt{D\eta}\cdot\sqrt{D\xi}} \le 1.$$

Число

$$\rho = \frac{|M(\eta - M\eta)(\xi - M\xi)|}{\sqrt{D\eta}\cdot\sqrt{D\xi}}$$

называют *коэффициентом корреляции* между ξ и η.

Отметим следующие его свойства:

1. $|\rho| \le 1$;
2. Если ξ и η независимы, то $\rho=0$.

   В самом деле, ввиду независимости имеем равенство $M[(\eta - M\eta)\cdot(\xi - M\xi)] = M(\eta - M\eta)\cdot M(\xi - M\xi) = (M\eta - M\eta)\cdot(M\xi - M\xi) = 0,$ и $\rho=0$.
3. $\rho = 1$ тогда и только тогда, когда между ξ и η существует линейная зависимость.

В самом деле, если $\rho = 1$, то дискриминант трехчлена f(a) равен нулю, и существует единственный корень уравнения $f(a) = 0$, обозначим его a_o. Тогда $f(a_0) = M[(\eta - M\eta) + a_0(\xi - M\xi)]^2 = 0$, выражение под знаком математического ожидания равно нулю, то есть $(\eta - M\eta) + a_0(\xi - M\xi) = 0$, или $\eta = M\eta + a_0(\xi - M\xi)$.

Обратно, если η линейно выражается через ξ:

$$\eta = a_0\xi + a_1,$$

то $M\eta = a_0M\xi + a_1$, $\eta - M\eta = a_0\xi + a_1 - (a_0M\xi + a_1) = a_0(\xi - M\xi)$, и $|\rho| = 1$.

Третье свойство коэффициента корреляции ρ дает возможность судить о качестве линейной модели регрессии.

Имея статистическую выборку $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ значений случайного вектора, вместо коэффициента корреляции используют его оценку

$$r = \frac{Sxy}{\sqrt{S_x^2}\cdot\sqrt{S_y^2}},$$

где $S_{xy} = \frac{1}{n}\cdot\sum_{k=1}^{n}(x_k - \bar{x})\cdot(y_k - \bar{y}) = \sum_{k=1}^{n}x_k\cdot y_k - \bar{x}\cdot\bar{y},$

$S_x^2 = \overline{x^2} - \left(\overline{x^2}\right) = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_y^2 = \overline{y^2} - \left(\overline{y^2}\right) = \frac{1}{n}\cdot\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Выборочный коэффициент корреляции r имеет свойства 1,3 коэффициента ρ, что позволяет использовать его как меру линейной связи между x и y.

Пример 34.1 По данным примера об агенте по продаже недвижимости (*пример 29.1*) находим

$$r = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{S_x^2}\cdot\sqrt{S_y^2}} = \frac{1683{,}86 - 18{,}17\cdot 88{,}84}{\sqrt{17{,}88}\cdot\sqrt{415{,}35}} = 0{,}805.$$

Как видим, значение r достаточно близко к единице, и выбор линейной модели оправдан. ◆

§35. Коэффициент детерминации

Определение 35.1. *Коэффициентом детерминации* называется квадрат коэффициента корреляции, ρ^2.

В статистических задачах употребляется *выборочный коэффициент детерминации*

$$r^2 = \frac{(\overline{xy} - \bar{x}\cdot\bar{y})^2}{S_x^2 \cdot S_y^2}.$$

С помощью простых преобразований эту формулу можно привести к эквивалентному виду:

$$r^2 = \frac{S_y^2 - [\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2]}{S_y^2}.$$

Для объяснения последнего выражения заметим, что отклонение y_i от $\bar{y}$ можно представить в виде

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

Графическая иллюстрация последнего равенства видна на рис. 35.1

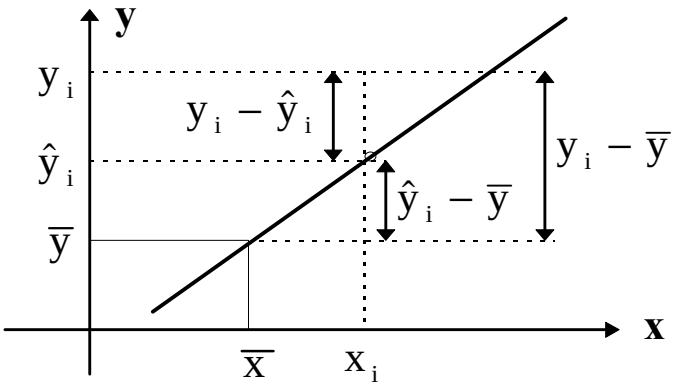


Рис. 35.1

Разность $\hat{y}_i - \bar{y}$ образуется в зависимости от разности $x_i - \bar{x}$, то есть вариация выходной переменной обуславливается входной, регулируемой переменной.

Вторая часть разности $y_i - \bar{y}$ есть разность между y_i и его оценкой, $\hat{y}_i$. Эта разность есть ошибка модели, в нее входит влияние неучтенных факторов (в примере с агентом по торговле недвижимостью это могут быть:

местоположение жилья, природные факторы, количество ванных комнат и спален в доме и т.п.).

Суммированием получаем следующее равенство:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2,$$

или, после перехода к средним,

$$S_y^2=RSS+\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

Учитывая равенство

$$RSS=S_y^2-\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2,$$

получаем для r^2 другое выражение:

$$r^2=\frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2]}{S_y^2},$$

то есть коэффициент детерминации измеряет ту часть вариации выходной переменной, которая вызывается изменением входной переменной.

Пример 35.1 По данным примера об агенте по продаже недвижимости (*пример 29.1*) получен $r = 0{,}805$, то есть $r^2 = 0{,}648$.

Это означает, что 64,8% изменчивости цены объясняется изменением общей площади жилья. Остаток - 35,2% изменчивости - объясняется неучтенными факторами.

Итак, общая площадь жилья есть превалирующий фактор в образовании его цены. ◆

§36. Заключительные замечания

В случае, когда коэффициент детерминации мал (степень этого определяется самим исследователем) возникает вопрос об улучшении качества модели за счет введения новых регулируемых переменных, приходя к линейной модели вида

$$y = a_0 + a_1x_1 + a_2x_2 + ... + a_kx_k,$$

где $x_1, x_2, ..., x_k$ - входные переменные, либо за счет усложнения модели, делая ее квадратичной, логарифмической, показательной, то есть выбирая ее в виде

$$y = a_0 + a_1x^2,$$

либо

$$y = a_0 \cdot \log_{a_1} x,$$

либо

$$y = a_0 \cdot x^{a_1},$$

и т.д.

Отыскание неизвестных параметров $a_0, a_1, ..., a_k$ производится с использованием метода наименьших квадратов, однако детальное изложение этих вопросов выходит за рамки настоящего курса.

Регрессионный и корреляционный анализ находит широкое применение при прогнозировании, при решении задач хозяйственного и внутрипроизводственного планирования. Практика показывает, что регрессионные уравнения - хорошие измерители связей между экономическими явлениями.

ПРИЛОЖЕНИЯ

Приложение 1

Таблица значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot \ell^{-\frac{x^2}{2}}$

x	0	1	2	3	4	5	6	7	8	9
0.0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3925	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3725	0.3712	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.3467	0.3448	0.3429	0.3410	0.3391	0.3372	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2897	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1.0	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1456	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.1040	0.1023	0.1006	0.0989	0.0973	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551
2.0	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0194	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0088	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.0061
2.9	0.0060	0.0058	0.0056	0.0055	0.0053	0.0051	0.0050	0.0048	0.0047	0.0046
3.0	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006
3.6	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004
3.7	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003
3.8	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002
3.9	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001

Приложение 2

Таблица значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x \ell^{-\frac{z^2}{2}} dz$

x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)
0.00	0.0000	0.5	0.1915	1.00	0.3413	1.50	0.4332	2.00	0.4772	2.50	0.4938
0.01	0.0040	0.51	0.1950	1.01	0.3438	1.51	0.4345	2.01	0.4778	2.51	0.4940
0.02	0.0080	0.52	0.1985	1.02	0.3461	1.52	0.4357	2.02	0.4783	2.52	0.4941
0.03	0.0120	0.53	0.2019	1.03	0.3485	1.53	0.4370	2.03	0.4788	2.53	0.4943
0.04	0.0160	0.54	0.2054	1.04	0.3508	1.54	0.4382	2.04	0.4793	2.54	0.4945
0.05	0.0199	0.55	0.2088	1.05	0.3531	1.55	0.4394	2.05	0.4798	2.55	0.4946
0.06	0.0239	0.56	0.2123	1.06	0.3554	1.56	0.4406	2.06	0.4803	2.56	0.4948
0.07	0.0279	0.57	0.2157	1.07	0.3577	1.57	0.4418	2.07	0.4808	2.57	0.4949
0.08	0.0319	0.58	0.2190	1.08	0.3599	1.58	0.4429	2.08	0.4812	2.58	0.4951
0.09	0.0359	0.59	0.2224	1.09	0.3621	1.59	0.4441	2.09	0.4817	2.59	0.4952
0.10	0.0398	0.6	0.2257	1.10	0.3643	1.60	0.4452	2.10	0.4821	2.60	0.4953
0.11	0.0438	0.61	0.2291	1.11	0.3665	1.61	0.4463	2.11	0.4826	2.61	0.4955
0.12	0.0478	0.62	0.2324	1.12	0.3686	1.62	0.4474	2.12	0.4830	2.62	0.4956
0.13	0.0517	0.63	0.2357	1.13	0.3708	1.63	0.4484	2.13	0.4834	2.63	0.4957
0.14	0.0557	0.64	0.2389	1.14	0.3729	1.64	0.4495	2.14	0.4838	2.64	0.4959
0.15	0.0596	0.65	0.2422	1.15	0.3749	1.65	0.4505	2.15	0.4842	2.65	0.4960
0.16	0.0636	0.66	0.2454	1.16	0.3770	1.66	0.4515	2.16	0.4846	2.66	0.4961
0.17	0.0675	0.67	0.2486	1.17	0.3790	1.67	0.4525	2.17	0.4850	2.67	0.4962
0.18	0.0714	0.68	0.2517	1.18	0.3810	1.68	0.4535	2.18	0.4854	2.68	0.4963
0.19	0.0753	0.69	0.2549	1.19	0.3830	1.69	0.4545	2.19	0.4857	2.70	0.4965
0.20	0.0793	0.7	0.2580	1.20	0.3849	1.70	0.4554	2.20	0.4861	2.72	0.4967
0.21	0.0832	0.71	0.2611	1.21	0.3869	1.71	0.4564	2.21	0.4864	2.74	0.4969
0.22	0.0871	0.72	0.2642	1.22	0.3888	1.72	0.4573	2.22	0.4868	2.76	0.4971
0.23	0.0910	0.73	0.2673	1.23	0.3907	1.73	0.4582	2.23	0.4871	2.78	0.4973
0.24	0.0948	0.74	0.2704	1.24	0.3925	1.74	0.4591	2.24	0.4875	2.80	0.4974
0.25	0.0987	0.75	0.2734	1.25	0.3944	1.75	0.4599	2.25	0.4878	2.82	0.4976
0.26	0.1026	0.76	0.2764	1.26	0.3962	1.76	0.4608	2.26	0.4881	2.84	0.4977
0.27	0.1064	0.77	0.2794	1.27	0.3980	1.77	0.4616	2.27	0.4884	2.86	0.4979
0.28	0.1103	0.78	0.2823	1.28	0.3997	1.78	0.4625	2.28	0.4887	2.88	0.4980
0.29	0.1141	0.79	0.2852	1.29	0.4015	1.79	0.4633	2.29	0.4890	2.90	0.4981
0.30	0.1179	0.8	0.2881	1.30	0.4032	1.80	0.4641	2.30	0.4893	2.92	0.4982
0.31	0.1217	0.81	0.2910	1.31	0.4049	1.81	0.4649	2.31	0.4896	2.94	0.4984
0.32	0.1255	0.82	0.2939	1.32	0.4066	1.82	0.4656	2.32	0.4898	2.96	0.4985
0.33	0.1293	0.83	0.2967	1.33	0.4082	1.83	0.4664	2.33	0.4901	2.98	0.4986
0.34	0.1331	0.84	0.2995	1.34	0.4099	1.84	0.4671	2.34	0.4904	3.00	0.4987
0.35	0.1368	0.85	0.3023	1.35	0.4115	1.85	0.4678	2.35	0.4906	3.04	0.4988
0.36	0.1406	0.86	0.3051	1.36	0.4131	1.86	0.4686	2.36	0.4909	3.06	0.4989
0.37	0.1443	0.87	0.3078	1.37	0.4147	1.87	0.4693	2.37	0.4911	3.08	0.4990
0.38	0.1480	0.88	0.3106	1.38	0.4162	1.88	0.4699	2.38	0.4913	3.08	0.4990
0.39	0.1517	0.89	0.3133	1.39	0.4177	1.89	0.4706	2.39	0.4916	3.12	0.4991
0.40	0.1554	0.9	0.3159	1.40	0.4192	1.90	0.4713	2.40	0.4918	3.16	0.4992
0.41	0.1591	0.91	0.3186	1.41	0.4207	1.91	0.4719	2.41	0.4920	3.20	0.4993
0.42	0.1628	0.92	0.3212	1.42	0.4222	1.92	0.4726	2.42	0.4922	3.26	0.4994
0.43	0.1664	0.93	0.3238	1.43	0.4236	1.93	0.4732	2.43	0.4925	3.32	0.4995
0.44	0.1700	0.94	0.3264	1.44	0.4251	1.94	0.4738	2.44	0.4927	3.40	0.4997
0.45	0.1736	0.95	0.3289	1.45	0.4265	1.95	0.4744	2.45	0.4929	3.60	0.4998
0.46	0.1772	0.96	0.3315	1.46	0.4279	1.96	0.4750	2.46	0.4931	3.80	0.49993
0.47	0.1808	0.97	0.3340	1.47	0.4292	1.97	0.4756	2.47	0.4932	4.00	0.499968
0.48	0.1844	0.98	0.3365	1.48	0.4306	1.98	0.4761	2.48	0.4934	4.50	0.499997
0.49	0.1879	0.99	0.3389	1.49	0.4319	1.99	0.4767	2.49	0.4936	5.00	0.4999997

Приложение 3

Таблица значений $\chi^2_{k\alpha}$, соответствующие вероятности $p = P\{\chi^2_k > \chi^2_{k,\alpha}\}$, где χ^2_k имеет χ^2- распределение с k степенями свободы

k	α										
	0,99	0,95	0,9	0,5	0,25	0,1	0,05	0,025	0,01	0,005	0,001
1	0	0	0,02	0,45	1,32	2,71	3,84	5,02	6,63	7,88	10,8
2	0,02	0,1	0,21	1,39	2,77	4,61	5,99	7,38	9,21	10,6	13,8
3	0,11	0,35	0,58	2,37	4,11	6,25	7,81	9,35	11,3	12,8	16,3
4	0,3	0,71	1,06	3,36	5,39	7,78	9,49	11,1	13,3	14,9	18,5
5	0,55	1,15	1,61	4,35	6,63	9,24	11,1	12,8	15,1	16,7	20,5
6	0,87	1,64	2,2	5,35	7,84	10,6	12,6	14,4	16,8	18,5	22,5
7	1,24	2,17	2,83	6,35	9,04	12	14,1	16	18,5	20,3	24,3
8	1,65	2,73	3,49	7,34	10,2	13,4	15,5	17,5	20,1	22	26,1
9	2,09	3,33	4,17	8,34	11,4	14,7	16,9	19	21,7	23,6	27,9
10	2,56	3,94	4,87	9,34	12,5	16	18,3	20,5	23,2	25,2	29,6
11	3,05	4,57	5,58	10,3	13,7	17,3	19,7	21,9	24,7	26,8	31,3
12	3,57	5,23	6,3	11,3	14,8	18,5	21	23,3	26,2	28,3	32,9
13	4,11	5,89	7,04	12,3	16	19,8	22,4	24,7	27,7	29,8	34,5
14	4,66	6,57	7,79	13,3	17,1	21,1	23,7	26,1	29,1	31,3	36,1
15	5,23	7,26	8,55	14,3	18,2	22,3	25	27,5	30,6	32,8	37,7
16	5,81	7,96	9,31	15,3	19,4	23,5	26,3	28,8	32	34,3	39,3
17	6,41	8,67	10,1	16,3	20,5	24,8	27,6	30,2	33,4	35,7	40,8
18	7,01	9,39	10,9	17,3	21,6	26	28,9	31,5	34,8	37,2	42,3
19	7,63	10,1	11,7	18,3	22,7	27,2	30,1	32,9	36,2	38,6	43,8
20	8,26	10,9	12,4	19,3	23,8	28,4	31,4	34,2	37,6	40	45,3
21	8,9	11,6	13,2	20,3	24,9	29,6	32,7	35,5	38,9	41,4	46,8
22	9,54	12,3	14	21,3	26	30,8	33,9	36,8	40,3	42,8	48,3
23	10,2	13,1	14,8	22,3	27,1	32	35,2	38,1	41,6	44,2	49,7
24	10,9	13,8	15,7	23,3	28,2	33,2	36,4	39,4	43	45,6	51,2
25	11,5	14,6	16,5	24,3	29,3	34,4	37,7	40,6	44,3	46,9	52,6
26	12,2	15,4	17,3	25,3	30,4	35,6	38,9	41,9	45,6	48,3	54,1
27	12,9	16,2	18,1	26,3	31,5	36,7	40,1	43,2	47	49,6	55,5
28	13,6	16,9	18,9	27,3	32,6	37,9	41,3	44,5	48,3	51	56,9
29	14,3	17,7	19,8	28,3	33,7	39,1	42,6	45,7	49,6	52,3	58,3
30	15	18,5	20,6	29,3	34,8	40,3	43,8	47	50,9	53,7	59,7

Приложение 4

Таблица значений $t_{k\beta}$, соответствующие вероятности $\beta = P\{|t_k| > t_{k,\beta}\}$,

где случайная величина t_k имеет распределение *Стьюдента* с k степенями свободы

k	β							
	0,2	0,1	0,05	0,02	0,01	0,005	0,002	0,001
1	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	1,886	2,92	4,303	6,965	9,925	14,09	22,33	31,6
3	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,61
5	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	1,44	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	1,397	1,86	2,306	2,896	3,355	3,833	4,501	5,041
9	1,383	1,833	2,262	2,821	3,25	3,69	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,428	3,93	4,318
13	1,35	1,771	2,16	2,65	3,012	3,372	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,14
15	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	1,337	1,746	2,12	2,583	2,921	3,252	3,686	4,015
17	1,333	1,74	2,11	2,567	2,898	3,222	3,646	3,965
18	1,33	1,734	2,101	2,552	2,878	3,197	3,61	3,922
19	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,85
21	1,323	1,721	2,08	2,518	2,831	3,135	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	1,319	1,714	2,069	2,5	2,807	3,104	3,485	3,768
24	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	1,316	1,708	2,06	2,485	2,787	3,078	3,45	3,725
26	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,689
28	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,66
30	1,31	1,697	2,042	2,457	2,75	3,03	3,385	3,646
40	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
60	1,296	1,671	2	2,39	2,66	2,915	3,232	3,46
120	1,289	1,658	1,98	2,358	2,617	2,86	3,16	3,373
10000	1,282	1,645	1,96	2,327	2,576	2,808	3,091	3,291

ЛИТЕРАТУРА

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика, Основы моделирования и первичная обработка данных. - М.: Финансы и статистика, 1983.

2. Вентцель Е.С., Овчаров Л.А. Теория вероятностей. - М.: Наука,1973.

3. Гмурман В.Е. Теория вероятностей и математическая статистика. - М.:
Высш. шк.,1972.

4. Гмурман В.Е. Руководство к решению задач по теории вероятностей и
математической статистике. - М.: Высш. шк.,1975.

5. Гурский Е.И. Теория вероятностей с элементами математической статистики. - М.: Высш. шк.,1971.

6. Колде Я.К. Практикум по теории вероятностей и математической статистике. - М.: Высш. шк.,1991.

7. Румшиский Л.З. Элементы теории вероятностей. - М.:
Наука,19730.

8. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. - М.: Финансы и статистика, 1983.

Учебное издание

Бандура Вячеслав Николаевич
Породников Виктор Дмитриевич

ТЕОРИЯ ВЕРОЯТНОСТЕЙ

И

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие